ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/284663966

Cervini C., Masperi M., Jouannaud M.-P., Scanu F. (2013). "Defining, modeling and piloting SELF, a new formative assessment test for foreign languages". In Language Testing in Euro...

Conference Paper · May 2013

CITATIONS
0

READS
99

4 authors, including:

Cristiana Cervini
University of Bologna; University Stendhal
15 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Monica Masperi
Université Grenoble Alpes
24 PUBLICATIONS **56** CITATIONS
SEE PROFILE

Marie-Pierre Jouannaud
Université Stendhal - Grenoble 3
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Recherches en intercompréhension View project

Project Fluence View project

All content following this page was uploaded by Cristiana Cervini on 25 November 2015.

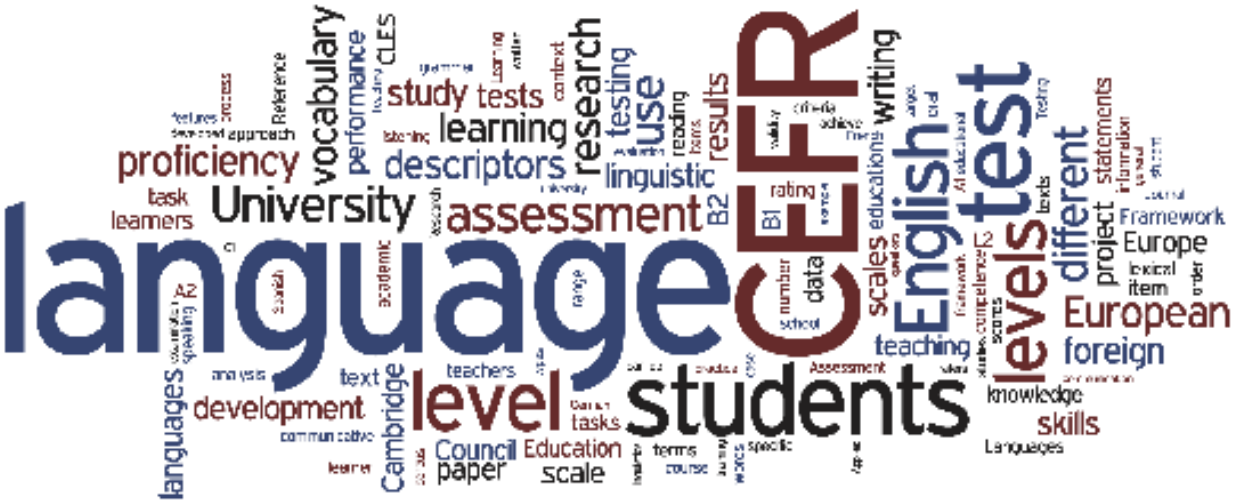
The user has requested enhancement of the downloaded file.

Proceedings

Language Testing in Europe: Time for a New Framework?

27-29 May 2013
Antwerp, Belgium

Jozef Colpaert
Mathea Simons
Ann Aerts
Margret Oberhofer



Linguapolis
Universiteit Antwerpen

Institute of Education
and Information Sciences

Universiteit
Antwerpen

Language Testing in Europe: Time for a New Framework?

Proceedings

University of Antwerp

27 – 29 May 2013

Composed by Jozef Colpaert, Mathea Simons, Ann Aerts, Margret Oberhofer

Jozef Colpaert, Mathea Simons, Ann Aerts, Margret Oberhofer (editors)

Proceedings, 2013, "Language Testing in Europe: Time for a New Framework?", Antwerp: University of Antwerp.

Cover: Nieuwe Media Dienst, University of Antwerp

ISBN 9789057284106
EAN : 9789057284106

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen of op enige manier, zonder voorafgaandelijke schriftelijke toestemming van de uitgever.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system
of transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher.

Uitgave en verspreiding:

Universiteit Antwerpen
Prinsstraat 13
2000 Antwerpen
www.ua.ac.be

FOREWORD

This volume contains the presentations from our Second International Conference on Language Testing. The first conference, which was organised in 1997, brought together more than 100 participants from 12 countries around the theme 'Language Testing and HRM´.

This Second International Conference unites more than 150 practitioners, policymakers and researchers from 26 countries. The theme, 'Language Testing in Europe: Time for a new Framework?' arose from an urgent need to respond to concrete issues associated with the use of the Common European Framework of Reference for Languages (CEFR) in language testing.

- *Competence and performance*
  What is the link between 'can do' performance statements and areas of linguistic knowledge? To what extent can or should the levels be made more explicit in terms of required vocabulary and grammar?

- *Degree of difficulty of the levels*
  How can we make sure that our examinations are measuring at the CEFR levels we claim they are? What evidence do we have to support our claims?

- *Test purpose*
  Why are we testing? What kind of decisions will be made on the basis of information collected via the test? What will be the consequences of these decisions?

- *Practicality*
  How do we link our tests to the CEFR? How practical, applicable and operational is the CEFR for concrete language testing situations?

Lyle Bachman (Professor Emeritus at the University of California), Etienne Devaux (Screening methodologist at SELOR, the Belgian public personnel selection and certification agency), Jan Hulstijn (Professor at the University of Amsterdam) and Waldemar Martyniuk (Executive Director of the European Centre for Modern Languages of the Council of Europe), will give keynote presentations in order to enhance the discussions from their highly specific areas of expertise.

Glyn Jones has been selected to receive the award for the Selected Plenary Presentation.

This volume contains 49 full papers in alphabetical order. All contributions demonstrate a remarkable variety in background, approach and style. They will provide the foundation for three days of intense debate on language testing and the CEFR.

One of the objectives of this conference is to formulate a clear memorandum of understanding. This document will reflect the opinion of each participant. To this end, we conducted a pre-conference survey (the results of which will be presented on the first day). We have also arranged various types of interaction with the audience, including a voting moment, an online survey, discussion groups, a panel and, most importantly, many coffee and food breaks.

We hope that you will enjoy this volume and that you will keep it as a memorable souvenir of a remarkable milestone in the history of language testing in Europe.

Prof Jozef Colpaert & Prof Mathea Simons

CONFERENCE ORGANISATION

Chair
Jozef Colpaert
jozef.colpaert@ua.ac.be
Institute of Education and Information Sciences – Linguapolis – University of Antwerp

Conference manager
Ann Aerts
ann.aerts@ua.ac.be
Linguapolis – University of Antwerp

Local organising committee
Jozef Colpaert
Mathea Simons
Ann Aerts
Margret Oberhofer

Scientific committee
Jozef Colpaert
Mathea Simons
Margret Oberhofer
Charles Alderson
Lyle Bachman
Jan Hulstijn

CONTENTS

KEYNOTE SPEAKERS

Lyle F. Bachman

University of California, Los Angeles, The United States

lfb@humnet.ucla.edu

How do Different Intended Uses and Different Views of Language Impact Language Assessment Practice?

Abstract

The use for which an assessment is intended is generally regarded as the most important consideration in its design and development. Similarly, defining the construct (the area, component, or aspect of language ability) we want to measure is widely considered to be a critical decision in the process of developing a language assessment. In practice, language assessments are used for a wide range of uses, or decisions, and historically, test developers have drawn on a variety of theoretical frameworks of language to define the construct to be measured.

For any particular test, but especially for large-scale, high-stakes tests, the test developers and test users are required, by current professional standards, to provide evidence to support their claims about the intended interpretations of test scores and about the intended uses of these interpretations to make decisions. However, in an increasingly global "market" of language testing, there is increasing pressure, on both test developers and test users, to find ways of "linking" different tests to a common conceptual framework of language use. What is at issue is that these tests may have been developed for very different uses, for different populations of test takers, and may be informed by very different views of the construct to be measured. In this environment, the demand for portability and transferability of interpretations often overrides fundamental concerns for reliability, validity, and fairness.

The primary purpose of "linking" different tests to each other or to a common standard is to enable test users to interpret and use the results of the two tests in the same way. Using two tests "in the same way" requires that the two tests measure similar constructs, that the decisions to be made are similar, and that the consequences of these decisions are similar. In my view, many current linking activities do not provide adequate justification for claims about these basic similarities. Given the pervasiveness of such claims and practice, I think it is imperative for us, as a profession, to address some very fundamental issues about the nature and justification of "linking" different language tests to a common standard.

In this presentation I will begin with a brief overview of the different language frameworks that have informed large-scale language tests in the past half century. I will then use an assessment use argument (Bachman & Palmer, 2010) to analyze the ways in which two different approaches to defining language differ in terms of the claims they make about score-based interpretations. I will then discuss the different uses for which tests based on these two ways of defining language might be most appropriate. Finally, I will return to the issue of the difficulty of "linking" tests based on different types of language frameworks and intended for different kinds of decisions to a common framework of language.

References

Bachman, L. F. & Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford: Oxford University Press.

Etienne Devaux

SELOR, Brussel, Belgium

etienne.devaux@selor.be

Confronting the CEFR to L2 Certification Purposes : Added Value and Methodological Limitations

Abstract

As an institutional personnel selection and certification agency, SELOR has a tradition of L2 certification in the Belgian context. Assessing the receptive and productive second language skills of public personnel belongs to our core missions.

In the past few years, we have adapted our tests and assessment methods to the real-life professional context and chose the CEFR as our reference. We worked with an academic experts panel for theoretical issues and involved experienced raters to operationalize decisions. We undertook the revision of the existing methodology and chose a validation model. We gathered in-depth information (qualitative feedback and quantitative studies) to identify satisfactory areas and aspects amenable to improvement. Our work revolved around the main axes of this model: the 'context validity' and 'theory-based validity' axes helped us revise our expectations and test specifications; we used the 'scoring validity' axis to revise our assessment tools and the rater training components. All new methods and contents were duly pretested and validated.

The discussion will focus on the added value and limitations of the CEFR, which we thought might be of interest to other organizations. We found in the CEFR useful common concepts to work in a multilingual context. We experienced that the CEFR is not a self-contained ready-to-use framework and that organizations need specialized partners to use it sensibly. We noticed that the construct choices we made in terms of competence and domains could jeopardize the assumed comparability of CEFR proficiency levels. We observed that the CEFR provides limited input about assessment methods and heterogeneous descriptors and that developing assessment tools requires great care and considerable investments. Using the CEFR for validation purposes and finding data fit for an external criterion validation may also be challenging. This discussion could open perspectives for further studies.

Jan H. Hulstijn

University of Amsterdam, Amsterdam, The Netherlands

j.h.hulstijn@uva.nl

Natural Tensions between Theory and Practice in the Common European Framework of Reference for Languages (CEFR)

Abstract

The CEFR (Council of Europe, 2011) represents a brave and moderately successful attempt to cater for the interests of stakeholders in the field of language education (learners, curriculum planners, schools, teachers, employers, local and national authorities), while taking account of theories of language use and language acquisition and the empirical research supporting these theories. However, in all walks of human life, there is a natural tension between practice and theory, simply because they serve different purposes. It is therefore impossible to base any framework of language teaching and assessment completely on theory and research in the language sciences. This does not mean, however, that language-acquisition theory and research are irrelevant for the CEFR. In this presentation, I will present a model of language proficiency in native and non-native speakers, and some hypotheses derived from it (Hulstijn, 2011; Hulstijn, in progress), proposed as both sufficiently plausible and sufficiently implausibe to deserve to be empirically falsified. The model, embedded in a usage-based approach to language acquisition, distinguishes between basic language cognition (BLC) and higher (or extended) language cognition (HLC). BLC is the language, used in the aural/oral modes, which all native speakers have in common; HLC concerns all other language knowledge and use. Despite the fact that this model, like most models and theories in science, does not render the observed phenomena correctly, I will propose that there is room to use it cautiously in discussions on the CEFR. The main danger currently threatening the CEFR is the diversity of interpretations of its levels, leading to unwanted differences in assessment practices within and across languages. To combat this threat, I will propose that the Vocabulary Range scale (CEFR, 2001, p. 112) be additionally defined in terms of vocabulary-size numbers and that vocabulary tests form part of language-proficiency exams.

References

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. Language Assessment Quarterly, 8, 229-249.

Waldemar Martyniuk

European Centre for Modern Languages of the Council of Europe, Graz, Austria

waldemar.martyniuk@ecml.at

The Council of Europe's Common European Framework of Reference for Languages (CEFR): a 2013 Summary of Developments

Abstract

The Council of Europe's Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) was developed by a Council of Europe international working party between 1993 and 1996 with a view to promote transparency and coherence in language learning and teaching in Europe. After a pilot scheme, it was officially published in English and French in 2001, proclaimed the European Year of Languages by the Council of Europe jointly with the European Union. The CEFR has since been translated into additional 38 languages and quickly turned to be one of the most influential publications of the last decade in the field of language learning, teaching and especially language testing in Europe and elsewhere. In a steadily growing number of countries, the CEFR has become a powerful instrument for shaping language education policies. The task of relating language policies, language curricula, teacher education and training, textbook and course design and content, examinations and certification systems to the CEFR has been undertaken by a considerable number of public and private stakeholders in Europe and beyond.

Since its publication in 2001, the CEFR has grown to become a core element of an extensive set of materials, a toolkit for different target groups. In my contribution, I intend to summarise the developments around the CEFR by referring to the many related recommendations, tools and instruments that have been developed by the Council of Europe over the last years and made available to the users of the Framework – as a starting point for the discussion on what may be next to happen for the CEFR, in Europe and beyond.

SELECTED PLENARY

Glyn Jones

Pearson, London, United Kingdom

glyn.jones@pearson.com

Developing a CEFR-Aligned Test from Scratch: a Case Study

Bio data

Glyn Jones worked as a teacher of English as a Foreign Language before becoming a specialist in Computer Assisted Language Learning (CALL) and self-access learning. Since 2001 he has been involved in language test development. He currently works as a senior researcher for Pearson in London.

Abstract

This paper is concerned with the genesis of a high stakes test of General English. The test was developed with the intention that each of its six levels should be aligned to the corresponding level of the CEFR, A1 to C2. Procedures recommended for the "Specification" stages of alignment in the Council of Europe's Manual (Council of Europe, 2009) were followed. The drafting of test specifications was informed by CEFR descriptors, as was the formulation of scoring criteria for speaking and writing.
However, various challenges were encountered in this process. Most notably:

- As has been pointed out by Alderson et al (2004) among others, the descriptors are far from complete in their coverage, from one level to the next, of language activities. The higher levels (C1 and C2) are especially underspecified in this respect.
- Even where descriptors do suggest appropriate assessment tasks (e.g "Can understand short simple letters" – *Reading Correspondence* at A2) they do not help with design decisions such as the linguistic features of texts to be used for comprehension testing, or the criteria for assessing productive tasks – decisions which are often critical in setting the level of difficulty of the task.

The presenter will relate how the test development team endeavoured to meet these challenges by

- Formulating item writer guidelines according to a schema which aims to specify how CEFR descriptors apply to critical aspects of task design, such as grammatical complexity or choice of distractors
- Applying a similar schema to the formulation of scoring rubrics

Short paper

Introduction

The Pearson Test of English General (PTE General) is a suite of six examinations in General English developed and operated by Pearson Language Testing. The test is taken on prescribed dates in approved test centres in a wide range of countries. It is assessed by means of two modules: a paper based test covering the skills of listening, reading and writing; and a speaking test in the form of a face to face interview.

The test was originally developed by the University of London Schools Examinations Board (ULSEB) and launched in 1985 under the title Certificate of Attainment in English.

At that time it was offered at five levels called simply "Level 1" to "Level 5" respectively. A sixth level was later added below Level 1. This was called "Level A1" to reflect the fact that it was designed to assess at CEFR Level A1.
The test was acquired by Pearson in 2003. In the meantime the name of the test had been changed to London Test of English (LTE).

In 2006 a thorough review of the test was initiated with a view to deciding whether a revision was necessary, and to determine the scope of any such revision. The review proceeded by

- Eliciting the views of external stakeholders by means of questionnaires administered to test centres, item writers and examiners, and by interviewing focus groups of candidates
- Critical evaluation by internal staff, including test developers, marketing staff and country representatives.

Proposals for change were reviewed by a panel of external experts, the Technical Advisory Group (TAG).

As a result of the review it was decided to institute a revision of the test. The main terms of reference for the revision were that

- The new test should be designed, from the outset, to be aligned to the CEFR rather than being aligned by post hoc mapping, with each of the six levels to be aligned to the corresponding CEFR level, A1 to C2.
- Items should be defined more rigorously than in the previous specifications so as to ensure greater similarity between parallel forms.

The basic format and coverage of the test were kept: i.e. the test was to remain a test of general English, offered at six levels, and assessing all four skills through two modules at each level.

This paper is concerned with the first of the above objectives, aligning to the CEFR, and will briefly relate the steps taken to this end and highlight some of the issues encountered.

Inclusion of CEFR descriptors in test specification
The specifications for the new version of the test contain a detailed definition of each item type. Each definition begins with a brief description of the item and a statement of assessment objectives. This is followed by a framed text box in which the CEFR descriptors are listed which relate most closely to the objectives. After this come detailed practical stipulations such as the types of text that can be used and their word limits and, finally, two or more sample items.

The reason for including CEFR descriptors in this way, and for giving them such prominence, was to give item writers an indication of what test takers are expected to be able to do at the respective level.

Familiarisation
Item writers and raters were given CEFR familiarisation training using procedures recommended in the Manual (Council of Europe, 2009), principally:

- Sorting of descriptors (on separate slips of paper) into scale order
- Identifying and discussing terms and phrases in the descriptors which serve to distinguish between adjacent levels
- Self-assessment in a second language using CEFR criteria
- Judging of benchmark samples of learner production (written and spoken)

Scoring of constructed response (speaking and writing) items
It was decided that the most direct way of linking the assessment of speaking and listening to the CEFR levels was to use actual CEFR descriptors as scoring rubrics.

Speaking tasks are rated according to six criteria, three of which – Range, Accuracy, Fluency and Interaction – are taken from the Qualitative aspects of spoken language use (Council of Europe, 2001:28f). A fifth criterion is taken from the scale for Phonological Control. The source of the sixth criterion varies according to task type: Sustained Monologue for a simple long turn; Turn Taking for a discussion task; Thematic Development for a long turn based on a visual stimulus; Sociolinguistic Appropriateness for a role play task. At a given level, the descriptors used for each of these criteria are taken verbatim from the corresponding CEFR scale. For example, in the criteria for PTE General Level 1 (the test which is designed to assess at CEFR level A2) the descriptor for Accuracy is "Uses some simple structures correctly, but still systematically makes basic mistakes", exactly as in the CEFR.

A similar approach was adopted for writing. Four generic criteria, applied to all tasks, are Range (from General Linguistic Range), Accuracy (from Grammatical Accuracy), Coherence and Cohesion (from Coherence), and Orthographic Control. A fifth, task specific criterion is taken either from Overall Written Interaction or Overall Written Production, depending on the task type.

For both skills, raters are instructed to award a score between 1 and 5 for each of the criteria. A score of 3 should be awarded if the performance meets the relevant descriptor (i.e. it is at the level). A score of 1 indicates a performance that is clearly below the level (i.e. it is at the next level down, or even lower). A score of 5 indicates a performance which is clearly above the level (i.e. it is at the next level up or even higher). Scores of 2 or 4 are awarded for borderline performances. This scheme is applied at all levels except at Level 5 where only the scores 1, 2 and 3 are awarded; 4 and 5 cannot logically be used as there is no higher level.

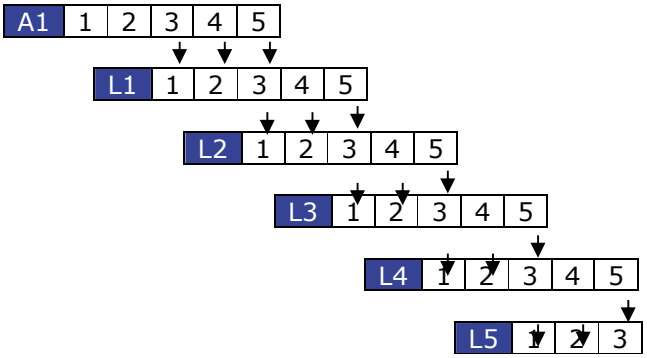


Figure 1: Overlapping marking scale for PTE General writing and speaking

This system ensures that scores from the different levels of the test can be mapped to the same underlying scale. This can be seen in Figure 1. A score of 4 (borderline with the next level up) at Level 2, for example, is equivalent to a score of 2 (borderline with the next level down) at Level 3.

Missing descriptors
In order to realise the above scheme in practice some judicious selection and editing were necessary. The distribution of CEFR scales between the four skills is very uneven; there are far more descriptors for speaking than for any of the other skills. The five generic scales adopted for assessment of speaking are all explicitly associated with this skill and four of them are grouped together in a single grid entitled Qualitative aspects of

spoken language use. However there is no equivalent grid for writing. The generic scales that have been adopted for this purpose are included in the CEFR publication under the heading "Linguistic competences", where only one of them, Orthographic Control, is explicitly linked to writing. The other three, by implication, may be associated with speaking or writing. In fact they are mostly identical in wording to the corresponding descriptors in the Qualitative aspects of spoken language use. This is not the only instance in which descriptors belonging to different scales are actually very similar or even identical.

Some scales needed to be supplemented, as in several instances the corresponding CEFR scale has no descriptor at some levels, or else states (at C2) "As C1". To plug these gaps we either

- borrowed a descriptor from another related scale. The descriptor for Turn Taking at C2 is taken from the highest level of the ELTDU scale (ELTDU, 1976); the descriptor for Phonological Control at C2 is an extract from the scoring rubrics for PTE Academic, a test that has been aligned to the CEFR independently.
- drafted a descriptor based on that for an adjacent level but with changes made to the wording to adjust the level of difficulty. The descriptor for Thematic Development at A1 is Can describe something using isolated words or simple phrases. This was derived from the A2 descriptor Can tell a story or describe something in a simple list of points, by positing a speaker who does not yet possess sufficient resources to tell a story or even to enumerate points in the form of a coherent list.

The scoring rubrics in practice

When, during piloting, raters started to use the scoring rubrics, they reported that they needed more guidance in applying the criteria. It was not so much a matter of the descriptors being insufficiently precise or detailed, although this could be the case, as of wondering what to accept as evidence that the descriptors had been met in the context of the task.

For example, the speaking test includes a short role play. At Level 2 (B1) this is of two minutes' duration and is rated for Sociolinguistic Appropriateness, among other criteria. The CEFR descriptor for this scale at B1 is:

Sociolinguistic Appropriateness
Can perform and respond to a wide range of language functions, using their most common exponents in a neutral register.

The following additional guidance is given as an indication of how "a wide range of language functions" is to be interpreted in this context.

[The test taker] may be required to perform the following functions and respond to them: requesting, offering, suggesting, thanking, rejecting, apologizing or congratulating.

Item writer guidelines

When item writers came to use the test specifications they found in practice that very often the CEFR descriptors were not sufficiently informative, by themselves, to enable them to set the degree of difficulty of an item appropriately for the level. Some reported that if they were able to write items at the appropriate degree of difficulty this was because they had internalised the construct through training (and in many case through long experience as a language teacher) and not because they had the relevant descriptors in front of them. It was therefore decided to develop item specific guidelines to help writers to align their items more precisely to the desired difficulty level. The aim was not to lend more precision to the CEFR descriptors themselves, but to attempt to

determine what the descriptors imply in relation to a given item type with all its attributes and constraints. In other words the question which the guidelines set out to answer is not "What can a learner at level X do?" but "What are the properties of a test item of this type such that a learner at level X can answer it?"

The procedure used in formulating the guidelines was as follows. For each item type at each level:

1. Determine which language ability(ies) the item is designed to test.
2. Determine which CEFR descriptors are applicable. To a large extent this has already been done as the CEFR descriptors were consulted in the process of drafting the test specifications in the first place, andsome of them are indeed reproduced there. For the purpose of the guidelines, however, the net was cast wider, to include any descriptors which might be helpful, not only the most relevant ones.
3. Identify, in the wording of those descriptors, the key terms which serve to pinpoint the level and to distinguish it from the levels above and below.
4. Simplify: in practice it turns out that similar or identical terms often occur in descriptors of different scales at the same level. These duplicates or near-duplicates are merged.
5. For each such term, decide what its implications are for each component of the item type in question. The list of components, in this sense, varies with item type. For a listening comprehension item, for example, the components are a listening passage, which has both textual and acoustic properties, and a task (what the test taker is required to do in relation to the passage). The task may be further analysed into subcomponents such as the stem and options of multiple choice items.
6. Edit the whole into a set of stipulations, again leaving out any that would otherwise be repetitious.

The key step in this process is step 5, determining practical implications of the terms. This is done by common sense judgment, drawing on the expertise and experience of test writers and teachers. As such it is open to the criticism that such judgments are ultimately subjective (though amenable to discussion). On the other hand, the way in which the judgments are arrived at is such as to provide a clear audit trail. If any of the guidelines are found to be unworkable or misleading in practice they can be traced back through the process to the original wording of the CEFR, and they can be reworked if necessary.

A worked example

Table 1 shows a partial view of the development of guidelines for a sample item type, in this case a reading comprehension item at B2. Having determined which CEFR scales are relevant (2nd row), the B2 descriptors in those scales are scanned for key terminology (3rd row, highlighted). (For the sake of brevity only two terms are identified in the example. In fact the terms "quickly identify" and "deciding whether close study is worthwhile" are also glossed.)

For each of these of these terms a statement is formulated as to what the term implies in relation to some property of the item (rows 4 and 5).

Item type	B2 section 5 Short text with a single multiple choice gapfill
Which CEF scale(s) are applicable?	• Overall reading comprehension • Reading for orientation • Reading for information and argument
Which terms need to be glossed?	B2: Reading for orientation Can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile
Implications of identify the content an relevance	The task should require the test taker to identify the topic of the text with precision, e.g. not just an ad for a holiday but for an adventure holiday.
Implications of a wide range of professional topics	Texts can be work related but accessible to the general reader; either texts to inform general readers about technical matters (e.g. information leaflets) or generic work related texts, e.g. about office procedures, document processing, line management etc.

Table 1: partial schema of development of item writer guidelines for a reading comprehension item type at B2

Once this process has been completed for all the key terms, the resulting statements are edited, where appropriate, so as to produce a set of stipulations for each relevant aspect of the item, in this case the text and the task ("audio", "picture" and "distractors" are other possible headings). These are arranged in order for each item type, with the original CEFR descriptors reproduced in an adjacent column for reference. Table 2 shows an extract from the finished product with the stipulations derived in Table 1 highlighted.

section	CEFR	B2 guidance notes
4	**B2 Overall reading comprehension** Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms. **B2 Reading for information and argument** Can obtain information, ideas and opinions from highly specialised sources within his/her field. Can understand articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints. **B2 Reading for orientation** Can quickly identify the content and relevance of news items, articles and reports on a wide range of professional topics, deciding whether closer study is worthwhile.	**Text** The text type may be any that a typical language user is likely to encounter in real life, including professional and academic situations The lexis in the text should be accessible to an educated general reader. The text should not contain highly colloquial or idiomatic expressions Texts can be work related but accessible to the general reader; either texts to inform general readers about technical matters (eg information leaflets) or generic work related texts, eg about office procedures, document processing, line management etc. **Task** The task should require the test taker to identify the topic of the text with precision, e.g. not just an ad for a holiday but for an adventure holiday The task should be designed to assess understanding of the purpose or main message of the text, or familiarity with the formal linguistic features of the genre, including stylistic features, register and appropriate vocabulary.

Table 2: extract from completed guidance

Conclusion

The CEFR is not itself a set of test specifications, of course. However, its authors do maintain that the Framework can be used "for the specification of the content of tests and examinations" (Council of Europe, 2001:19). This is what we have endeavoured to do in the course of developing the new version of PTE General. In the process we have encountered issues that have required elaboration of the test documentation beyond what can be derived directly from the CEFR. Some of these issues are due to structural features of the framework that have been commented on by others (notably by Alderson et al., 2004), such as missing descriptors, imprecise or inconsistent terminology, and duplications. Others arise from the fact that the people who carry the responsibility for operationalizing test specifications in practice – item writers and raters – often ask for more detailed guidance than the Framework itself can provide. We have endeavoured to address these issues in ways that are effective and transparent.

References

Alderson, J. C. , Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). The development of specifications for item development and classification within The Common European Framework of Reference for Languages: Learning, Teaching,

Assessment: Reading and Listening: Final report of The Dutch CEF Construct Project. Lancaster.

Council of Europe. (2001). Common European Framework for Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe. (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): a manual. Strasbourg: Language Policy Division.

ELTDU. (1976). Stages of attainment scale and test battery: general information. Oxford: Oxford University Press.

PAPER PRESENTATIONS

Ene Alas & Suliko Liiv

Tallinn University, Tallinn, Estonia

enealas@tlu.ee - liiv@tlu.ee

Training Interviewers and Raters for the National Examination in the English Language – the Estonian Experience

Bio data

Ene Alas is a lecturer at Tallinn University. Her research interests are language testing, test development and evaluation and teacher education

Suliko Liivi is a professor at Tallinn University. Her research interests are contrastive linguistics, intercultural communication, language policy and language testing.

Abstract

As of 2014, the national examination in the English language in Estonia will attempt to measure students' language proficiency within two levels on the CEFR scale (B1 and B2) instead of focusing on only B2, as has been the practice so far. The blueprint of the new examination is expected to be available in spring 2013. Changing the concept of the national examination has meant, among other things, altering all sections of the national examination, developing a new interviewer script for the speaking section, designing new marking scales for the subjectively marked sections – writing and speaking – and training interviewers to use the interviewer script as well as training raters to work with the new marking scales. The presentation will briefly concentrate on the challenges posed by the development of the marking scales and relating them to CEFR. The speakers' main focus, however, is another aspect of quality control - the practical aspect of training teachers of English in Estonia, who act as interviewers and raters within the framework of the national examination in the English language, to use the script and the marking scales reliably. A proposal will be made for a training sequence to reach that end.

Short paper

Now that high-stakes language proficiency test development has been firmly established in Estonia, research is needed regarding the model for assessing speaking within the framework of the national examination in the English language. Research will bring forth what the examiners' and assessors' attitude to the model is, how closely it is followed and what further training needs there are. Our concern in this connection is a lack of local examples of Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR) level performances. A database is needed of audio/videotaped local performances of speaking tasks that would serve as benchmarks for the assessors. Such benchmarking alongside with defining the borderline cases would serve as a step in the direction of linking the national examination in the English language to the CEFR. The procedure needs empirical data as well as theoretical discussion of the proposed practice. Another important concern in that context is the role of the cultural background of the interviewer/ assessor/ interviewee, i.e., to what extent the cultural background of the participant may appear as a variable in the evaluation process and how this variable affects the benchmarking process. On the basis of this research, pre- and in-service training programmes need to be developed for the qualification and re-qualification of the national examination speaking test interviewers

and assessors. The content of such programmes and the process of developing them would be the other subjects of the current project. Methods of the current research: questionnaire-studies and interviews with interviewers/ assessors, content analysis of assessment interviews, modelling assessor behaviour using multivariate statistical analysis, linking quantitative and qualitative data (mixed methods approach) to identify factors leading to biased assessment. The research would be conducted in cooperation with S/A Innove (former the National Examination and Qualification Centre).

Research results will be published in academic articles, introduced at international research events and in a summary project document. A new lecture course will be developed for degree students.

The current research is partly supported by Estonian Science Foundation Grant No. ETF 9037.

Practicality

How do we link our tests to the CEFR? How practical, applicable and operational is the CEFR for concrete language testing situations?

The national examination in the English language has been operational since 1997. It is administered at the end of gymnasium, is one of the optional exams but one which is very frequently chosen by students (about 7000 to 9000 students choose it every year). The exam is not officially linked to the CEFR, but is claimed to be testing if the students taking it have reached CEFR B2 level of language proficiency. Students are tested in writing, listening, reading, language structures (these four parts make up the written paper of the test) and speaking (tested on a separate day). The claim made so far that the test is attempting to be a B2 level test is based on the results of the steps taken that should facilitate linkage with the framework: test developers are expected to demonstrate and are given training in familiarization with the level descriptors of the framework; the examination content and task types are audited to see if they meet the expectations on B2 level; training is provided to test developers, assessors and interviewers to ensure and maintain standard behaviour during examination administration; standard setting is conducted in terms of establishing cut scores and borderline cases. What is missing, however, is a 'principled set of procedures and techniques that provides support in what is a technically complicated and demanding process' (Noijons et al 19) that has been adopted for the current testing system that would provide a systematic approach to the matter. With the introduction of a new national examination that is intended to test the same population on two levels (B1/B2), the need for a systematic approach is indispensible if informed decisions about the language proficiency level demonstrated during the exam are to be issued to the examinees.

One of the problems related to linking the exam to the CEFR is the size and quality of the sample required from the candidates to determine their particular level. And if the test attempts to assess students on 2 levels, what does it mean in terms of the number and content of tasks? Do they need to be doubled, how much is enough? Given the limited amount of time that the students have for test completion, can the evidence provided during the test be enough to make a decision about their proficiency level. Thus auditing the content of the test is problematic.

The situation is further complicated by procedural problems during the examination administration: our research shows that there is limited consistency among interviewers during oral interviewers, only 20% of the trained interviewers fully followed the interviewer script (Alas, 162). Thus relating the exam with a variation in standards might be a further issue.

References

Alas, E. (2010). The English Language National Examination Validity Defined by Its Oral Proficiency Interview Interlocutor Behaviour. Diss. Tallinn: Tallinn UP.

Noijons, J., Bérešova, J., Bretton, G.& Szabó, G. (Ed.). (2011). Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR). Highlights from the Manual. Council of Europe.

Marian Amengual Pizarro & Jesús García Laborda

University of the Balearic Islands, Palma de Mallorca, Spain
University of Alcalá de Henares, Madrid, Spain

marian.amengual@uib.es - jesus.garcialaborda@uah.es

Investigating Teachers' Opinion on the Feasibility and Intended Washback Effect of a High-stakes Oral English Test in Spain

Bio data

Marian Amengual-Pizarro currently holds the post of lecturer in the Department of Spanish, Modern and Classical Philology (English Philology) at the University of the Balearic Islands. Her research has been mainly in the area of testing. She has been coordinating the English Test in the Spanish University Admission Examination at the University of the Balearic Islands (UIB) since 2003. Her other main research area is the teaching of English as a foreign language. She has been the 'Language Teaching' Panel coordinator at AESLA (Spanish Association of Applied Linguistics) for eight years (2004-2008). She is also the current secretary of AEDEAN Association (Spanish Association of Anglo-American Studies, and member of the editorial board of the journal "e-resla". She has published in national and international refereed journals on applied linguistics and education.

Jesús García Laborda is an associate professor at Universidad de Alcala (Madrid, Spain). Dr Garcia Laborda has a PhD in English Philology and an EdD in Language Education. His current research covers many areas of computer implementations for language learning and testing along with ESP and teacher training especially for the Spanish University Entrance Examination or the implications of implementing such test in teacher training along with more traditional approaches to teacher education and their development of both cognitive and computer skills. His publications include papers in Computers & Education, British Journal of Educational Technology or Educational Technology and Society.

Abstract

Research into the influence of tests on teaching and learning referred to as 'washback' or 'backwash' in the education literature has been extensive over the past decades. The majority of washback studies have emphasised the negative consequences of tests, especially high-stakes examinations, on different areas of the curriculum. More recently, however, high-stakes tests (used for making important decisions which affect people's futures) have been employed to reform instruction and achieve positive washback (Weir 1990; Spolsky, 1996; Norris 2009). Indeed, many different countries in the world have introduced various types of high-stake tests with the aim of improving education and support good practice (Alderson and Wall, 1993; Cheng, 2004; Qi, 2007). The future inclusion of an oral English sub-test in the Spanish University Admission Examination (academic year 2013-2014) is seen as an attempt to improve the level of spoken English among Spanish undergraduates and promote positive washback.

This study investigates the opinion of 13 secondary teachers (out of a total of 15) who participated in the implementation of the pilot oral test conducted in Majorca (Balearic Islands). The teachers evaluated a total of 175 secondary students in May 2012. Results, collected from a questionnaire, show that teachers hold positive views on the organization, structure and design of the new oral test. Furthermore, the majority of

them believe that the oral test will affect teachers' methodology and increase the amount of time devoted to the practice of students' oral skills in class (Amengual, 2009), although some concerns are raised over gains obtained due to coaching for the examination. Findings also reveal teachers' concerns associated with the need to receive some training courses in the use of the rating scales to ensure rater inter-reliability. Finally, most teachers question the feasibility of developing this test due to the current economic situation of the country.

Short paper

This paper aims at investigating the feasibility and potential washback effect of the new English oral sub-test in the Spanish University Admission Examination (SUAE) to be put into effect in 2014. The Spanish University Admission Examination (SUAE) is a high-stakes public examination taken annually by millions of students at the end of their secondary education in order to enter a Spanish university. The current English Test (ET) format across the majority of Spanish universities fails to evaluate important communicative abilities of the students since it mainly concentrates on candidates' reading and writing abilities and, therefore, it is not considered a valid measure of communicative language ability. Moreover, it is generally believed that preparing students for the ET has negative consequences for the practice of oral communication since most of the class time is devoted to the teaching of skills featured in the ET. The future inclusion of an oral sub-test in the design of the ET as proposed by the Spanish education authorities is seen as an attempt to achieve beneficial washback and meet the ever-increasing demand for more communicative English tests.

This investigation, therefore, deals with two of the main conference topics: 'Test purpose' and 'practicality'. The primary purpose of the new English oral sub-test is to promote positive washback. Research into the influence of tests on teaching and learning referred to as 'washback' or 'backwash' in the education literature has been extensive over the past decades. The majority of washback studies have emphasised the negative consequences of tests, especially high-stakes examinations, on different areas of the curriculum. More recently, however, high-stakes tests (used for making important decisions which affect people's futures) have been employed to reform instruction and achieve positive washback (Weir 1990; Spolsky, 1996; Norris 2009). Indeed, many different countries in the world have introduced various types of high-stake tests with the aim of improving education and support good practice (Alderson and Wall, 1993; Cheng, 2004; Qi, 2007). The future inclusion of an oral English sub-test in the Spanish University Admission Examination (academic year 2014) is seen as an attempt to improve the level of spoken English among Spanish undergraduates (since it is believed it will encourage teachers to increase the amount of time and attention given to speaking practice in class), and promote positive washback. To this end, the study investigates the opinion of 13 secondary teachers (out of a total of 15) who participated in the implementation of the pilot oral test conducted in Majorca (Balearic Islands). The teachers evaluated a total of 175 secondary students in May 2012. Results, collected from a questionnaire, show that teachers hold positive views on the organization, structure and design of the new oral test. Furthermore, the majority of them believe that the oral test will affect teachers' methodology and increase the amount of time devoted to the practice of students' oral skills in class (Amengual, 2009), although some concerns are raised over gains obtained due to coaching for the examination.

This investigation also examines the degree of difficulty some teachers had in assigning levels as well as 'test practicality'. The criteria presented as CEFR Table 3 for "qualitative aspects of spoken language use" (CEFR, 2001: 7) was used to develop the rating instrument of the new English oral sub-test, which was defined in terms of the following five analytical criteria: Range, Accuracy, Fluency, Interaction and Coherence. CEFR global oral assessment scales were also implemented. Teachers in the study reported having difficulties in weighing the components and assigning levels. Since teachers received no

training in the use of scales, that result came as no surprise. Indeed, findings in the study reveal teachers' concerns associated with the need to receive some training courses in the use of the rating scales to ensure rater inter-reliability. As far as 'test practicality' is concerned, most of the teachers question the feasibility of developing the new English oral sub-test due to the costs it entails in terms of human and economic resources, further aggravated by the current economic situation of the country. Nevertheless, all teachers believe the new ET design will exert a positive influence or washback effect on English language instruction and better respond to the real communicative needs of students in contemporary society.

References

Alderson, J. C. & Wall, D. (1993). Does Washback exist? Applied Linguistics 14 (2), 115-129.

Amengual-Pizarro, M. (2009). Does the English Test in the Spanish University Entrance Examination influence the teaching of English? English Studies *90* (5), 582-598.

Council of Europe. (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

Cheng, L. (2004). The Washback Effect of a Public Examination Change on Teachers' Perceptions towards their Classroom Teaching. In L. Cheng & Y. Watanabe (Ed.), Washback in language Testing (pp. 147-170). Mahwah, NJ: Lawrence Erlbaum.

Norris, J. M. (2009). Understanding and Improving Language Education through Program Evaluation: Introduction to the special Issue. Language Teaching Research 13 (1), 7-13.

Qi, L. (2007). Is Testing an Efficient Agent for Pedagogical Change? Examining the intended Washback of the Writing Task in a High-stakes English Test in China. Assessment in Education, 14 (1), 51-74.

Spolsky, B. (1996). The Examination-classroom Backwash Cycle: Some Historical Cases. In D. Nunan, R. Berry & V. Berry (Ed.), Bringing about Change in Language Education (pp. 55-66). Hong Kong: The University of Hong Kong, Department of Curriculum Studies.

Weir, C. (1990). Communicative Language Testing. Hemel Hempstead: Prentice Hall.

Bernadette Brouwers

Australian Council for Educational Research, Melbourne, Australia

bernadette.brouwers@acer.edu.au

The Assessment of Language Competence – Moving Forward

Bio data

Bernadette Brouwers holds a Masters Degree in Applied Linguistics from the University of Melbourne and is Project Director for the Assessment of Language Competence (ALC) Certificates at the Australian Council for Educational Research (ACER). She has worked as a language teacher for many years and has held leadership roles in professional learning and curriculum development in a range of secondary schools. Bernadette has also held key positions in language curriculum planning and development at a systemic level. She held the role of Manager of the Languages other than English Key Learning Area for the Victorian Curriculum and Assessment Authority (VCAA), where she led the developement of curriculum for 44 languages at senior secondary level.

In her current role with the ALC she has expanded the project to include a test of Spanish and is further developing the tests to ensure that they more effectively reflect international developments in language learning and assessment.

Abstract

Australian schools do not participate in national benchmark testing for second languages. In 1990, the Australian Council for Educational Research (ACER), in collaboration with the Australian Bicentennial Multicultural Foundation and the University of Melbourne's Language Testing Research Centre (LTRC), developed the Assessment of Language Competence (ALC) tests in listening and reading comprehension at three levels and in six languages (Chinese, French, German, Italian, Japanese and Modern Greek). The multiple choice tests are offered to schools annually.

The ALC is widely used by government and independent primary and secondary schools across all Australian states and territories as well as in New Zealand and into SE Asia and the Pacific. The test is not mandated and is used by schools for various purposes. These include motivating students (through the awarding of ALC certificates), formally comparing individual school performance with other schools (via the ALC School Report) and internal monitoring of language programs at the school level (through analysis of ALC School Report data within and across languages).

ALC tests are developed by language specific experts contracted to ACER. Test specifications were originally developed in consultation with the LTRC and formed the basis of guidelines for writing panels. Tests were extensively trialed and piloted and detailed teacher feedback informed the review of early testing.

Item writers attend a training workshop and are guided by a detailed Writers' Manual. Analysis of the ALC achievement descriptors forms part of the training workshop. These are reviewed annually to ensure they reflect curriculum and assessment developments at national and international levels. Descriptors draw on contexts, topics, communicative functions, text-types and item intents originally informed by the Australian Language Levels Guidelines.

This paper will provide an overview of the ALC item development process and will explore a process for more formally benchmarking and linking the ALC to the CEFR.

Short paper

Introduction
This paper will outline the existing Assessment of Language Competence (ALC) second language test development process and present an initial process for reviewing the tests in order to more formally align with and benchmark them against the CEFR. It will also raise some final questions for consideration in this undertaking.

Background
Over the past three decades in Australia there has been continuous recognition of the need to address the question of what and how well students learn languages (Scarino et al, 2011, p. xi). The ALC was established by the Australian Council for Educational Research (ACER) partly in response to this need and has now been offered to schools for over 20 years. It was originally referenced against the Australian Language Levels Guidelines. While not mandated, the ALL Guidelines formed a coherent framework for the design, implementation, assessment and evaluation of second language programs in Australia and provided a sound foundation for the ALC.

ALC tests were originally offered in Chinese, French, German, Italian, Japanese and Modern Greek. Modern Greek was replaced by, the increasingly popular Indonesian and, more recently, Spanish has been added to the suite in recognition of its importance as a world language and in response to requests from schools.

While schools use the test results for various purposes: motivating students (through the awarding of ALC certificates), formally comparing individual school performance with other schools (via the ALC School Report) and internal monitoring of language programs at the school level (through analysis of ALC School Report data within and across languages), recent inquiries from schools indicate that there is a strong interest in a more internationally recognised test. Changing demands of educational systems with increasing external accountability requirements (Hill, 2012, p. 42) at national and international levels have also prompted the ALC project to review its content and processes to ensure that they remain responsive to the needs of client schools. To these ends, the ALC has begun to explore the possibility of linking to the CEFR.

ALC item development process: strengths and limitations
A team of 19 ALC item writers develop 294 listening and 208 reading multiple-choice items per annual test cycle. Writers attend an annual writer training session and are guided by a detailed manual which includes timelines, descriptors, planning grids, annotated sample items and global checklists. These are followed more carefully by some writers than others and this is an area which needs to be addressed in order to ensure stronger consistency across tests.

Original target language items are developed online via a wiki and are intensively vetted by expert staff within ACER at the initial stages. Items are then downloaded into a draft test form and sent to external second vetters who are invited to provide feedback on the appropriateness of the levels and internal consistency as well as specific advice on individual items in need of revision.

Once the first draft tests have been finalised, they are sent to external proofreaders with expertise in the relevant language. Tests are also sent to an ACER expert proofreader for consistency of overall style and format. While the Project Director has ultimate responsibility for the final form of all tests and ACER has ownership of the items, writers have active input into all stages of the review process.

Tests are then trialled in a small number of classes across a range of schools and qualitative feedback is sought on areas such as the length of time need to complete each test, the difficulty level of questions relative to the certificate level of the test, the clarity of instructions, and the overall quality/suitability of the test materials.

The project undergoes a process of annual review based on an analysis of test data as well as on specific and general survey feedback from teachers whose students have sat the tests. This combined information is provided to the writers as a guide for the next cycle of item development and to update the Writers' Manual.

Apart from a set of common items across the certificate levels within each language test, the ALC is limited by the lack of a formal mechanism for tightening the links within and across test levels/languages. It is also limited in that it is only benchmarked against the cohort sitting a test for a particular language/certificate level in any given year.

Proposed process to begin linking the ALC to the CEFR
The process of reviewing the ALC and relating the tests to the CEFR will need to be overseen by a panel of language and testing experts both from within and external to ACER, in light of the various stakeholder interests. Given the varying degrees of familiarity with the CEFR, there is likely to be a need to provide familiarisation activities to ensure that all panel members have a strong, shared understanding of the CEFR.

As a guide, the project proposes to consider learnings from the Language Testing and the CEFR: Time for a New Framework? conference and to essentially follow the five inter-related sets of procedures outlined in the Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual.

To ensure the manageability of the process, two languages are likely to be initially selected for linking and it would be expected that this process could be replicated for additional languages.

Questions for resolution
A number of questions arise from a proposed review of the ALC to enable linking to the CEFR and include the following:

How will link items be developed across languages?
Can the ALC be linked to the CEFR and still reflect the national developments in the Australian curriculum for languages?
What is the best approach to training writers who are geographically dispersed?

Summary
Both the development processes of the existing ALC tests and the proposed linking to the CEFR, point to a multi-layered undertaking. The previous questions add an additional level of complexity.

To be successful, this project will need to carefully manage the distinct but complimentary interests of both the longstanding ALC tests and the CEFR.

References

Council of Europe. (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual. Strasbourg: Council of Europe.

Hill, K. (2012). Classroom-Based Assessment in the School Foreign Language Classroom. Frankfurt: Peter Lang.

Scarino, A., Vale, D., McKay P. & Clark, J. (1988). Australian Language Levels Guidelines. (ALL) Canberra: Curriculum Development Centre.

Scarino, A., Elder, C., Iwashita, I., Kim, S.H., Kohler, M. & Scrimgeour, A. (2011). Student Achievement in Asian Languages Education. Part 1: Project Report to the Australian Government Department of Education, Employment and Workplace Relations.

Jasminka Buljan Culej

National Centre for External Evaluation of Education, Zagreb, Croatia

jbculej@ncvvo.hr

ESLC in Croatia: Relation between Years of Learning the Foreign Language and the CEFR Level Achieved

Bio data

Jasminka Buljan Culej graduated from the Department of Biology, Faculty of Sciences and Mathematics at the University of Zagreb in 1992. In 1995, she graduated from the Catholic Faculty of Theology at the same university. In 1999, she finished the Postgraduate Study in Toxicology at the University of Zagreb with the thesis "Significance of Placenta in the Assessment of Metal Exposure in Women", receiving an MSc. In 2004, she received a PhD from the University of Zagreb, with the topic of her doctoral dissertation being "Application of Bone morphogenetic protein (BMP) and mechanism of action in treating osteoporosis." From 1996 to 2005 she worked as a researcher at a number of institutions: the Institute for Medical Research and Occupational Health, Zagreb; Department of Human Reproduction, General Hospital "Sveti Duh", Zagreb; and the Laboratory for Mineralized Tissues at the School of Medicine, University of Zagreb. From 2005 to 2006 she had been employed at the Croatian Ministry of Science, Education and Sport as a consultant to the Minister's Cabinet. From 2006 onwards, she is the Head of the Research and Development Department of the National Centre for External Evaluation of Education in Zagreb, Croatia.

Jasminka has received the 31st European Calcified Tissue Symposium Young Investigatory Award in 2004 and the American Society of Bone and Mineral Research Young Investigator Award in 2005. During her studies and research career she attended numerous professional trainings pertaining to her research interests; most notably the Policy Planning and Analysis for Improving Quality in Education at the Graduate School of Education, Harvard, USA in 2008. She is the author in more than 40 journal articles, conference papers and chapters in books.

Jasminka held an invited lecture in the European Parliament, Brussels 2nd October 2012, during the seminar "Multilingual Europe: lessons from the European Survey on Language Competencies" on the topic: Results of the ESLC project in Croatia.

Abstract

Foreign language teaching in Croatia starts in the first grade of primary education. Learning one foreign language is compulsory for all students from the first grade. In the fourth grade, students have the opportunity to choose a second foreign language as an optional subject.

ESLC testing was conducted in primary schools on a representative sample of eighth grade students at ISCED level 2. Student sampling was done by SurveyLang experts on the basis of the list of all Croatian eighth graders currently learning foreign languages.
During the administration of the ESLC, a total of 1,109 (49.6%) students were tested in English, and 1,126 (50.4%) students were tested in the second target language, German.

The majority of participating students (80%) had been learning English for five to eight years, while approximately 18% of students had been learning the first target language since kindergarten.

Students tested in the first target language (English) generally achieve good results in Reading, which was expected. Level A2 and higher, which is considered to be attainable after eight years of foreign language learning, is achieved by 54% of students. Second target language achievements show that 72% of students achieve level A1 and higher in Reading.

When comparing achievements in Listening, the results for the first target language show that 71% of students achieve level A2 and higher. The results in second target language Listening show that 76% of students achieve A1 and higher, while expectations were not met by 24% of students.

In the first target language Writing, 75% of students reached A2 and higher and twenty-five percent of students score below the expected minimum level, out of which 5% achieve pre-A1. Results for second target language Writing show that only 13% of students do not achieve A1 and higher.

Short paper

In this paper, we present the relation between years of learning the foreign language and the CEFR level achieved, based on the European Survey on Language Competences (ESLC) in Croatia in the school year 2011/2012. First, we give a short overview of primary education and foreign language teaching in Croatia. Second, we present the results of the analysis of relation between years of learning a language and the CEFR level achieved on a Croatian sample for two foreign languages, English and German.

Foreign language teaching in Croatia
In Croatia, primary school represents the compulsory level of education whose function is to ensure that students gain a broad education.

Eight-year primary education in Croatia is compulsory and free for all children from the ages of six to fifteen. Children must be six years old by the end of March to begin school the following September. Even though the official policy is that students can begin school in the year when they turn six, children typically begin primary school at the age of seven because their parents feel they will benefit from being more mature.

Primary education consists of three segments:

1. compulsory primary education conducted in regular primary schools and special institutions for students with developmental difficulties;
2. arts education conducted in primary music and dance schools;
3. primary education of adults conducted in regular schools and specialized institutions.

Foreign language teaching in Croatia starts in the first grade of primary education. Learning one foreign language is compulsory for all students from the first grade, while in the fourth grade they have the opportunity to choose a second foreign language as an optional subject. For international reference, grades one to four correspond to ISCED 1, while grades five to six correspond to ISCED 2.

Since 2003, students in Croatia have started to learn their first foreign language from the first grade, and until the fourth grade they have two first foreign language lessons per week. From the fifth grade onwards, students have three first foreign language lessons

per week. This means that the total number of first foreign language lessons at ISCED1 and ISCED2 levels is 700.

For the time being, a second foreign language is an optional subject. Students have two second foreign language lessons per week from the fourth or fifth grade until the eighth grade. Therefore, the maximum number of foreign languages lessons is 350. According to CNES, the Syllabus for Primary Education and CEFR, it is estimated that students who finish ISCED 1 (grades one to four) can achieve A1 in the first foreign language, and that students who finish ISCED2 can achieve A2.

Second foreign language students, who start to learn foreign languages in the fourth grade, are able to achieve the A1+ level, i.e. higher than the preparatory level. This means that students achieve a level higher than A1 in language skills, but they do not achieve A2 due to the limited number of lessons, except in rare cases.

According to the data for primary schools in the school year 2010/2011, of the 472,250 students enrolled in primary schools, 10.6% learned English in the first grade, while 3,5% students learned German. Italian and French was also studied by a smaller number of first-graders. By the eighth grade, the final grade of primary education in Croatia, 14.5% students learned English, 17% learned German, 14,5% learned French, 16.3% learned Italian, and 15.3% learned Spanish.

In conclusion, most students in Croatian schools learn English as their first foreign language and German as their second foreign language.

Years of learning the foreign language and the CEFR level
ESLC testing was conducted in primary schools on a representative sample of eighth grade students at ISCED level 2. Student sampling was done by SurveyLang experts on the basis of the list of all Croatian eighth graders currently learning foreign languages. Stratification was based on school size and six regions. A total of 1,109 (49.6%) students were tested in English, and 1,126 (50.4%) students were tested in the second target language, German.

According to Croatian law, written parental consent is needed for the participation of underage children in surveys. In 0.3% of cases, parents did not give consent, while 0.36% of students were unable to participate in the survey for justified reasons. Only 0.06% of students declined to participate in the survey. The main survey sample included 3,625 students with a response rate of 92.2%. This confirms that students in Croatian primary schools are still not overwhelmed by testings and that they are motivated to participate in surveys.

Before presenting the data, we have to bear in mind that when determining the language competences of each student, they do not necessarily have the same level of competence in all language skills. A student may possess more fully developed receptive language skills than productive language skills; however, this depends on their individual interests and affinities, living environment, language teaching conditions and other factors. Therefore, students may achieve A1 in the area of productive skills after the fourth grade, or they may achieve A2 in the area of comprehension even before finishing the eighth grade.

Therefore, we will compare the results of students differing in the number of years studying the language in Reading, Listening and Writing both for the first target language (English) and the second target language (German). All comparisons are done on a descriptive level.

Comparison of years of learning and CEFR for English

The majority of participating students (80%) had been learning English for five to eight years, while approximately 18% of students had been learning the first target language since kindergarten. Therefore, those students had been learning the first target language for more than eight years.

In Reading, the abovementioned 18% of students possibly achieve better results. In total, 43% of them achieve B2, 19% achieve B1, 11% achieve A2 and 28% of students do not achieve A2. In comparison, eighty percent of students had been learning the first target language for five to eight years, but 50% of them do not reach A2. Seventeen percent of students do not achieve A1, while 32% of them are at A1. Therefore, 50% of students achieve A2 and higher: 13% achieve A2, 14% achieve B1 and 23% achieve B2.

The results for Listening indicate that students who started learning the foreign language at an early age are possibly more successful at the task. In total, 51% of students achieve B2, 22% achieve B1, 11% achieve A2 and 17% do not achieve A2. Eighty percent of students had been learning the first target language for five to eight years, but 31% of them do not reach A2. Nineteen percent of students achieve A1 and 16% achieve A2. In total, 69% of students achieve A2 or higher: 15% are at A2, 23% are at B1 and 31% are at B2.

If we compare the achievements of students who had been learning English since kindergarten with the achievements of students who started learning English in primary school, the biggest difference is visible at pre-A1 level, where the number of students who started learning English at an early age is two times smaller than the number of students who started learning their first foreign language in primary school. Furthermore, we can point out the difference between the two categories of students at B2: approximately 20% of students who started learning English at an early age achieve this level.

Writing tasks require the use of higher cognitive levels where students demonstrate active foreign language proficiency. As expected, the results are somewhat lower than the results in Reading and Listening. As in previous analyses, here we also see that students who started learning foreign languages at an early age are more successful at all levels of Writing.

Regarding the other skills, in the category of students learning foreign languages since the first grade of primary school, we note that the number of students at pre-A1 and B2 levels is lower and they are more evenly distributed in the remaining three categories. Twenty-three percent of students achieve A1, 31% achieve A2 and 34% achieve B1.

If we take a look at the results of the students who had been learning foreign languages for more than eight years, 88% of them achieve A2 and higher. Only 5% of students are at pre-A1 and 20% are at A1.

Comparison of years of learning and CEFR for German

The results of students tested in the second target language, German, are analyzed in a similar fashion as the results of students tested in English. However, since German is the second target language, we can expect a different clustering of students by years of learning.

The number of students who had been learning German for one to four years is 62 or 5.5%. Students who had been learning the second target language from five to six years are the largest group; consisting of 617 (55%) students. 399 (35.6%) students had been learning the target language from seven to eight years, while 43 (3,9%) had been learning German for more than eight years.

Considering the distribution of students mentioned above, we shall discuss only the results of the second category of students, those who had been learning the second target language from five to six years (Category 1), and the results of the third category of students, those who had been learning the second target language from seven to eight years (Category 2).

In the first category of students, 74% achieve A1 and higher in German Reading, while 50.6% of students achieve A1. The proportion of students who do not achieve the targeted level is 26% and refers to pre-A1, while 13% of students achieve A2. There is only a small proportion of students who demonstrate a high level of proficiency: 6% achieve B1 and 4% achieve B2.

In the second, somewhat smaller, category of students, 74% achieve level A or higher; 44% achieve A1, 15% achieve A2, 10% achieve B1 and 5% achieve B2. Just as in the previous category of students, 26% of students are at the level of pre-A1.

According to the results mentioned above, there are no potential differences in Reading achievement between students who had been learning the second target language for five to six years and those who had been learning it for two years longer.

If the sample was larger, we would probably be able to define the potential differences more clearly. Perhaps the cause of the similarity in achievement is the fact that the difference in years of learning the foreign language is too small to be significant according to the statistical method used in the analysis.

The results for second target language in Listening are as follows.

In the first category of students, 77% are at A1 or higher, while 47.4% are at A1. The proportion of students who do not achieve the targeted level is 22.6% and refers to pre-A1, while 14.6% of students achieve A2. There is only a small proportion of students who demonstrate a high level of proficiency: 9.8% achieve B1 and 5.6% achieve B2.

In the second, somewhat smaller, category of students, 75% achieve level A or higher. Thirty eight percent of students are at A1, 20% are at A2, 11% are at B1, 5% are at B2 and 25% are at pre-A1.

Almost no differences were found between the levels of achievement in second target language Listening and second target language Reading. There may be differences in the second category of students at B2; however, a further study would be necessary to examine this in more detail.

The results for the second target language in Writing are as follows.
In the first category of students, 83% achieve A1 and higher in German Writing, while 53% of students achieve A1. The proportion of students who do not achieve the targeted level is 17% and refers to pre-A1, while 21% of students achieve A2. There is only a small proportion of students who demonstrate a high level of proficiency: 7% achieve B1 and 2% achieve B2.

In the second, somewhat smaller, category of students, almost 79% achieve level A or higher; 48% achieve A1, 21% achieve A2, 9% achieve B1 and 0.5% achieve B2, while 21% of students are at pre-A1.

In general, according to the results analyzed above, it seems that there are no significant differences between the first and second category of students regarding the level of achievement in Listening and Reading at A1 and higher. Perhaps we could find possible differences that would indicate better achievement, especially at level A2 for the category

of students who had been learning the target language for five to six years. However, the achievement of B1 and B2 is possibly lower in both categories of students.

If we compare the results of Croatian students with the results of students in countries that delay the onset of compulsory foreign language education until fifth grade (the French and Flemish Community of Belgium, Bulgaria and Netherlands), it is evident that students from those countries achieve lower results. For instance, if we look at CEFR levels achieved in English reading and listening, the performance of students in Bulgaria and the French Community of Belgium is lower than the performance of Croatian students. However, early foreign language education may not be the only factor influencing the overall lower results of students in the aforementioned countries. Namely, students in the Netherlands and in the Flemish Community of Belgium are among the best in all English skills, and they start learning English at a later age (SurveyLang, 2012; p. 23-24).

References

Croatian Parliament. (2008). State pedagogical standards for primary education. Zagreb: Narodne Novine. Retrieved from http://narodne-novine.nn.hr/clanci/sluzbeni/339618.html

Ministry of Science, Education and Sports of the Republic of Croatia. (2010). National Curriculum Framework for pre-school education, general compulsory and secondary education. Retrieved from http://public.mzos.hr/Default.aspx?sec=2497

Ministry of Science, Education and Sports. (2006). Croatian national educational standard: Syllabus for primary education. Retrieved from http://public.mzos.hr/Default.aspx?sec=2197

SurveyLang. (2012). First European Survey on Language Competences: Final Report, Version 3.0. Brussels: European Commission.

Kris Buyse

Katholieke Universiteit Leuven, Leuven, Belgium

kris.buyse@arts.kuleuven.be

About the Impossibility of Assessing Speaking with Focus both on Form and Communicative Output

Bio data

Kris Buyse is associate professor at the KU Leuven Faculty of Arts (Applied Language Studies, Teacher Training and Leuven Language Institute). He teaches Spanish proficiency (writing, Economical Spanish, Medical Spanish) and Didactics of Spanish as a Foreign Language. He is responsible for the Spanish Teacher training program. He is also visiting professor at the Universidad Antonio de Nebrija (Madrid) and at the Universidad Nacional de Educación a Distancia (UNED, Madrid). He has published on the corpus based investigation of translation, lexicography, contrastive linguistics, LSP, CALL and Spanish as a foreign language (especially vocabulary acquisition, writing, pronunciation and assessment). He has led several projects on educational development on vocabulary learning, writing and language teaching and assessment. He is reviews editor and member of the editorial committee of ITL-Journal of applied linguistics.

Abstract

"Assessing speaking is not impossible, but difficult (…): teachers often focus narrowly on the development of grammatically accurate speech which may conflict with a learner's desire to communicate and be understood" (Luoma 2004). Nevertheless, in the last two decades, with communicative and task based approaches as mainstream language teaching methodologies, assessments is focusing on authentic communicative contexts in which not only knowledge, but also skills and attitudes are required (Keeves 1994; Parrondo Rodríguez 2004). A growing number of publications is offering us models and patterns in order to simulate this authentic reality in the classroom (such as Cabré and Gómez de Enterría 2006 for Spanish), but without inquiring into the criteria, scales and templates suitable for a flexible evaluation of performance assessment of language competence in this communicative context in combination with a focus on form, apart from the templates based on and limited to the Common Framework of European Reference for Languages (CEFR(L)), which in our analysis will turn out to be too rigid when assessing both communicative output and focus on form (see also CITO/SLO 2010), as well as regarding the combination of formative and summative evaluation and in a analytic/synthetic way, two other criteria in current assessment, apart from validity, reliability and transparency (Dochy and Gijbels 2010).

Therefore, after presenting briefly the state of the art on performance assessment and the objectives of current language teaching, we will evaluate a corpus of assessment templates with respect to the abovementioned criteria.

Based on the data of our analysis, we will finally propose flexible criteria and templates for a communicative assessment of oral language skills, allowing us to adapt the assessment to the demands of any language course without losing sight of the five criteria. So yes, assessing speaking with focus both on form and communicative output is possible.

Short paper

Introduction

According to Dochy and Gijbels (2010), a high quality assessment is based on three criteria, i.e. validity, reliability and transparency. In the last two decades, with communicative and task based approaches as mainstream language teaching methodologies, emphasis is not only laid on transparency and reliability —promoting the use of more transparent templates and scales in a formative assessment process—, but also on validity, aiming at assessments in authentic communicative contexts in which not only knowledge, but also skills and attitudes are required (Keeves 1994; Parrondo Rodríguez 2004). A growing number of publications is offering us models and patterns in order to simulate this reality in the classroom (such as Cabré and Gómez de Enterría 2006), but without inquiring into the criteria, scales and templates suitable for a flexible evaluation of performance assessment of language competence in this communicative context in combination with a focus on form, apart from the templates based on and limited to the Common Framework of European Reference for Languages (CEFR(L)), which in our analysis turn out to be too rigid when assessing both communicative output and focus on form (see Buyse 2012 & 2013).

Therefore, after presenting briefly the state of the art on performance assessment in general and the objectives and demands of current language teaching and based on the analysis of our corps of assessment templates in Buyse 2012 & 2013, we will propose flexible criteria and templates for a communicative assessment of oral language skills, allowing us to adapt the assessment to the objectives and demands of each language course without losing sight of transparency, validity, reliability and user-friendliness.

Our view on current language teaching and assessment

In the context of current teaching evaluation does not constitute a component after teaching, but forms an intrinsic component of teaching itself, and one of the most guiding ones —i.e. which allows the student to orient his learning—, providing that (i) all parties involved (teachers and students) have access to detailed information on progress and problems, and that (ii) the teacher not only directs knowledge transfer but also accompanies the student in his search for information and his construction of (linguistic) knowledge and communication skills.
http://www.microsofttranslator.com/bv.aspx?from=es&to=en&a=http%3A%2F%2F131.253.14.66%2Fbvsandbox.aspx%3F%26dl%3Den%26from%3Des%26to%3Den%23_ftn2
(Sluismans and Dochy 1998).
Therefore, the object of evaluation not only concerns knowledge but also skills and attitudes (Keeves 1994). Performance assessment evaluates to what extent the student is able to apply the acquired knowledge to new problems (Meyer 1992). Realistic or simulated exercises should be used in order to do so (Gipps 1994; Parrondo Rodríguez 2004).

The integration of evaluation in the process of teaching and learning has put into prominence the term assessment ("continuous monitoring and evaluation") for this type of evaluation, because it emphasizes amongst others the use of competences, the relevance of an authentic evaluation context —i.e. that simulates realities in which the knowledge, skills and attitudes under assessment will be needed—, the importance of giving feedback, the quality of the input for the apprentice (Dochy and Gijbels 2010).
Considering assessment a component of teaching also emphasizes the close relationship between the objectives of learning and assessment. The former stipulate today in most curricula (in secondary, higher or adult education) that students assume their responsibility in the learning process by taking initiatives, interacting and handling heuristic instruments (such as dictionaries, grammars, corpus, portals, bibliographic search engines, etc.) in order to select relevant information. However, according to Gijselaers (2007) current generations need triggers and rewards in order to do so. This is where formative evaluation[1] can play a prominent role, giving a reward for the good work

done so far by the student and at the same time a stimulus or trigger for continuing in the same direction or making changes according to the feedback. Formative assessment is considered to be essential today, not only for the student because it allows him to detect positive elements and problems, as well as to formulate objectives and concrete ways for improvement, but also for the teacher, who can give more guidance to the student (dependent on the time available), taking into account the level and progress of the group, as well as of the individuals (Van Iseghem 2010). According to Dochy and Gijbels (2010), the frequent use of formative assessment allows to increase, expand and direct the motivation of the student, as well as deepen learning while a merely summative evaluation risks " learning to the test", i.e.: that the primary objective of the student is passing the test/exam.

According to Buyse 2012 & 2013, the templates that are used to evaluate the extent to which the student has achieved the objectives are becoming more analytical, i.e.: that the final score is the sum of separate notes for different criteria, such as the ones listed in the CEFR: general linguistic competence, richness of vocabulary, grammatical accuracy, vocabulary range, pronunciation range... Global or synthetic evaluation[2] http://www.microsofttranslator.com/bv.aspx?from=es&to=en&a=http%3A%2F%2F131.253.14.66%2Fbvsandbox.aspx%3F%26dl%3Den%26from%3Des%26to%3Den%23_ftn6carries a greater risk of entailing a "halo effect", i.e.: that the (too much) positive or negative evaluation of a single criterion influences the overall evaluation (Dochy and Gijbels, 2010). On the other hand, the intuition of experienced evaluators can provide us with a valuable synthetic assessment in comparison with the rather atomistic overview of the analytic assessment.

Another feature of the context of current teaching is the high degree of transparency, validity and reliability[3] required for assessment in education, and a growing tendency among students to contest the decisions of the evaluator(s). Apart from assessing in a maximally transparent way and controlling the validity of the assessment, an appeal can be avoided by raising the level of reliability. Therefore, an evaluation template can be used, after having been developed and accepted by a team of teachers/evaluators is preferably made within a teaching group, using keys (i.e. explanatory note on how to handle the template) and a shortlist (i.e. a list of objectives which have to be reached; see Dochy and Gijbels 2010: 124-125). Furthermore, the quality of the assessment procedure and materials should be regularly evaluated too (Van Iseghem 2010; Robles Avila et al. 2006). A practical requirement is that the number of criteria in a template should be limited —in specialized literature the maximum of criteria varies between 5 and

What is the link between 'can do' performance statements and areas of linguistic knowledge? To what extent can or should the levels be made more explicit in terms of required vocabulary and grammar?

As already announced in the introduction, in our corpus we did not find any templates which, besides being sufficiently valid, reliable and transparent, adequately reflect the characteristics and demands we just described, i.e. (1) assessing at the same time and in a flexible way the communicative output and the linguistic accuracy, (2) combining the advantages of synthetic and analytical assessments, (3) being sufficiently flexible for use in any type of education and their corresponding objectives, (4) allowing to combine summative and formative assessment of knowledge, skills and attitudes, (5) turning into a sufficiently detailed and reliable instrument to give feedback to the student. Most templates are based on and limited to the Common Framework of European Reference for Languages (CEFR(L)), which allows to assess communicative output and linguistic accuracy at the same time, but whose flexibility in our view is minimal: every change in the objectives, level or education type requires the development of another template: CITO/SLO (2011: 23) points out that, as CEFR distinguishes between oral expression, interaction and understanding, each skill also requires its own evaluation. CEFR offers the possibility of an integrated assessment of language skills, but in this case "there must be understood that in the case of a poor student performance, it is difficult to find out what

may be the cause" (ibid.), reducing seriously the chances of feedback, while the templates themselves do not pay much attention to the combination of the formative and summative assessment, nor to the analytic/synthetic dichotomy. In addition, according to the same authors, the evaluation through the CEFR constitutes a difficult task, because there are 13 criteria of which a selection has to be made. On the other hand, "what happens with a test in which the teacher wants to assess skills and linguistic components at the same time? Can he claim that his evaluation is consistent with the vision of the CEFR? It is important to note that the emphasis is on the evaluation of skills as reading, listening, speaking, talking and writing. If this is not the case, it becomes difficult to argue that the evaluation has been based on the CEFR" (CITO/SLO 2011: 23). See among others O' Sullivan (2011: 16-18, 104) for other types of concerns concerning the CEFR descriptors.

In the light of these deficiencies, we developed a series of flexible templates, i.e.: adaptable to all types and levels of courses according to their respective objectives, and capable of evaluating at the same time the communicative competence and the linguistic accuracy depending on the importance given to each of these criteria in the objectives of the course. The solution proposed here is to split this up into two columns, one for the evaluation of form (or "Text": what the student knows) and another for the communicative output (or "Task": what the student does), respectively (see Table 1). The criteria within each column are adaptable to the objectives of each course, although it seems indispensable that in addition to the expression also understanding is evaluated. The criterion of the "communicative output" evaluates the degree of communicative adequacy of the oral expression and interaction ("know-how"), while the criteria vocabulary, grammar and pragmatics evaluate to what extent adequate linguistic forms are used ("knowledge"), i.e.: the proper lexicon, suitable grammatical structures, in addition to the registry, coherence and cohesion, the compensation and interaction structures, etc. In this way the number of criteria is reduced to six (two times three) and there is no need to develop separate templates for oral expression and oral interaction, respectively. The relative weight of each column and each criterion depends on the objectives of the course and level: the importance of "text" could be higher in a language career than in a LSP course, for example, since the formers objective is to train language specialists.

Another fundamental aspect is the formative aspect, including for example positive elements per criterion and a reference to the learning process of the student, with a bonus system, accompanied or not by penalty system for negative attitude and progress (see Table 2).
In this template the evaluator has the possibility of evaluating in two phases: first he notes down positive and negative observations per category, but without rating them, just giving a global, intuitive note (in this case between AAA and E); in a second phase (for example, during the time that the next student receives instructions or later on when the evaluator takes a break) the performance is rated according to the sum of the scores for all categories. In case of a major discrepancy between the overall assessment and the analytic one, possible causes should be searched for. The answers on the surveys during our workshops and the feedback of the evaluators who are using these templates show that this procedure by phases generates more confidence in the reliability of the evaluation and is said to be easier to use, because during an oral test it is difficult to note down remarks and score appropriately at the same time. In this way scoring is more reliable and the risk of a discrepancy between the overall assessment and the analytic one is lower, according to the testimonies received.

Depending on the preferences of the evaluator, it is possible to add more "layers of evaluation": in Table 3 sub-criteria have been added to the categories, and in Table 4 assessment scales are added, which makes it for many assessors too difficult to use by the large amount of information on the template.

Although they may be adapted to the type of course, its level, the vision of the evaluator and the institute, the templates need regular updates, taking into account the problems and discussions among colleagues. A tool that allows different evaluators to check to what extent his evaluation corresponds to the one of other colleagues, is WebCef[4], where one can compare his own assessment of a recorded test with the ones of other colleagues.

Finally, in order to increase reliability, it is essential that each version of a template is accompanied by a key that contains the objectives and the subject of the course, the scoring system, its descriptors, the series of activities and selection system, as well as lists of typical problems by level.

Conclusions and discussion

In conclusion, the analysis of a corpus of oral proficiency assessment templates suggests that a more flexible template is needed, i.e. adaptable to all types and levels of courses according to their respective objectives. Especially the relationships between communicative goals and its evaluation, between knowledge and skills, and between linguistic accuracy and the communicative output require a different treatment.

The templates that have been presented in the previous section are meant to meet these requirements. They also include formative criteria in order to give the greatest possible amount of comments to students on their performance during the oral exam or test. In this sense they are rather assessment than "evaluation" instruments. This term originates in the Latin word "asidere", which means "sitting next to someone" and highlights the learning process (Dochy and Gijbels 2010). Therefore, the metaphor that teachers can keep in mind when evaluating, is that it is better to sit next to the student that in front of him.

Notes

[1] "Formative evaluation is generally any evaluation that takes place before or during a project's implementation with the aim of improving the project's design and performance. Formative evaluation complements summative evaluation and is essential for trying to understand why a program works or doesn't, and what other factors (internal and external) are at work during a project's life. Formative evaluation does require time and money and this may be a barrier to undertaking it, but it should be viewed as a valuable investment that improves the likelihood of achieving a successful outcome through better program design."
(http://evaluationtoolbox.net.au/index.php?option=com_content&view=article&id=24&Itemid=125, accessed 26-11-2012)

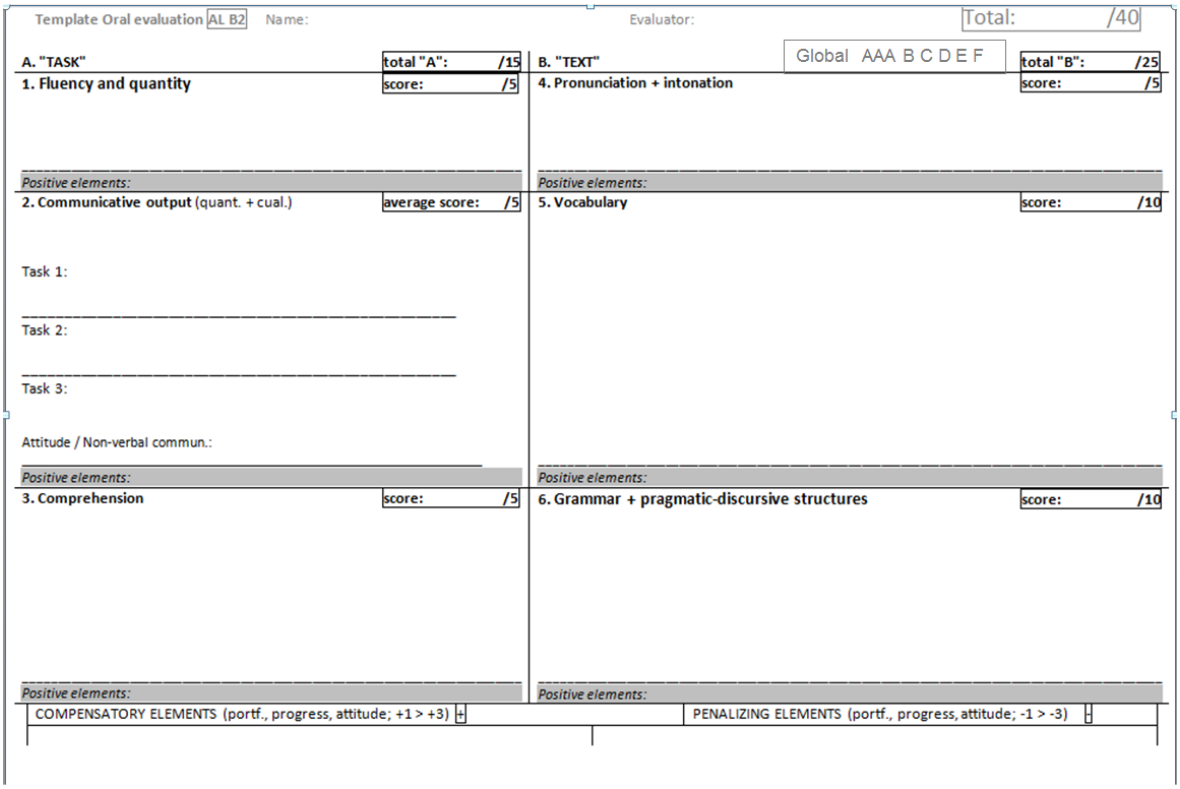
[2] "Summative evaluation looks at the impact of an intervention on the target group. This type of evaluation is arguably what is considered most often as 'evaluation' by project staff and funding bodies- that is, finding out what the project achieved. Summative evaluation can take place during the project implementation, but is most often undertaken at the end of a project. As such, summative evaluation can also be referred to as ex-post evaluation (meaning after the event)."
(http://evaluationtoolbox.net.au/index.php?option=com_content&view=article&id=40&Itemid=126, accessed on 26-11-2012)

[3] "Validity is a concept that deals with the framework used. It can be said that a test or a evaluation procedure is valid to the extent that it can be shown that what is assessed ("construct") is what, in the context in question, should be evaluated and that the information obtained is an exact representation of the linguistic students or candidates who performed the examination.
(…) Reliability, on the other hand, is a technical term. It is basically the degree in which the same order of the candidates in terms of the qualifications obtained in two different calls (real or simulated) of the same assessment test is repeated." (Instituto Cervantes, 2002: 177)

[4] www.webcef.eu

task		text	
communicative output		pronunciation	
fluency		vocabulary	
understanding		grammar & pragmatics	
	total		total
	%		%
overall total			

Table 1. "Task" and "text".



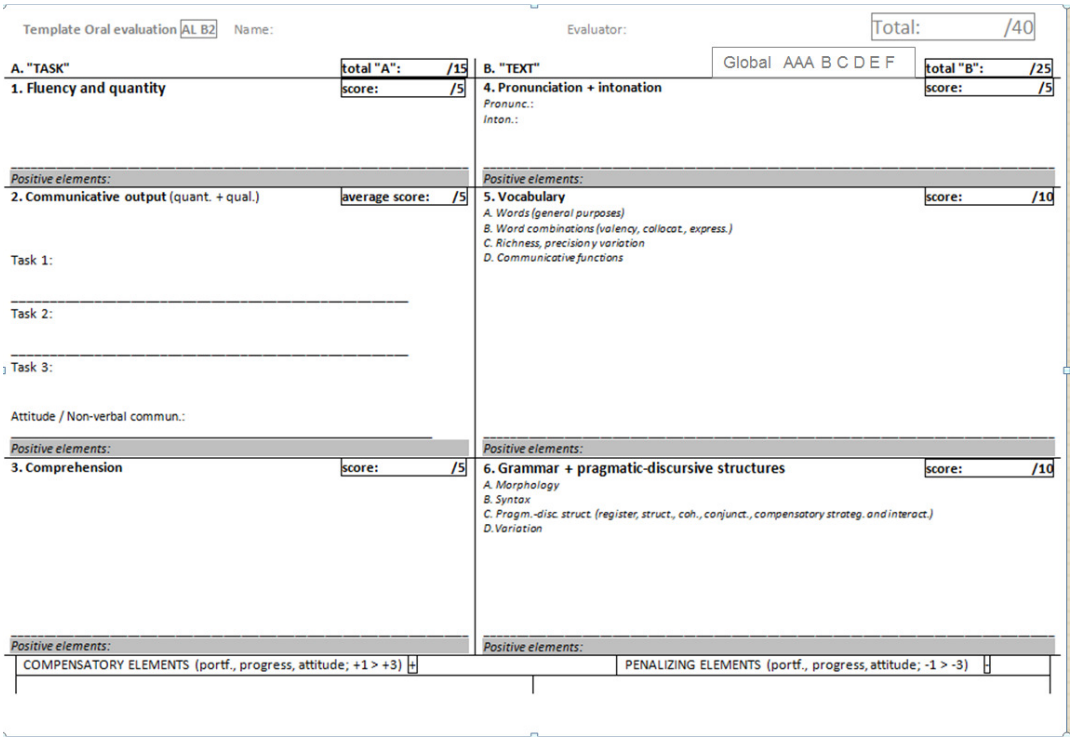
Template Oral evaluation AL B2 Name: Evaluator: Total: /40

Global AAA B C D E F

A. "TASK"	total "A": /15	B. "TEXT"	total "B": /25
1. Fluency and quantity	score: /5	4. Pronunciation + intonation	score: /5
Positive elements:		Positive elements:	
2. Communicative output (quant. + cual.)	average score: /5	5. Vocabulary	score: /10
Task 1:			
Task 2:			
Task 3:			
Attitude / Non-verbal commun.:			
Positive elements:		Positive elements:	
3. Comprehension	score: /5	6. Grammar + pragmatic-discursive structures	score: /10
Positive elements:		Positive elements:	

COMPENSATORY ELEMENTS (portf., progress, attitude; +1 > +3) +	PENALIZING ELEMENTS (portf., progress, attitude; -1 > -3) -

Table 2. Template for oral competencies in the course of translation and interpreting, level B2.

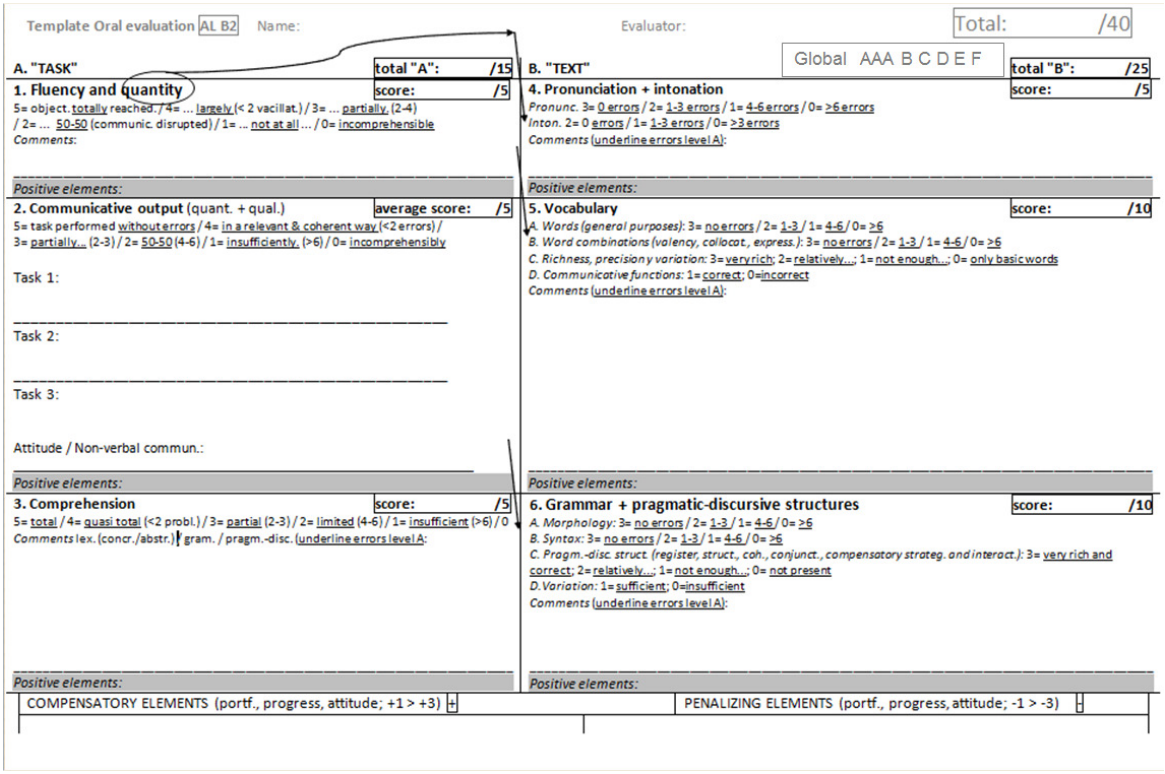


Template Oral evaluation AL B2 Name: Evaluator: Total: /40

A. "TASK"	total "A": /15	B. "TEXT"	Global AAA B C D E F	total "B": /25
1. Fluency and quantity	score: /5	4. Pronunciation + intonation		score: /5
		Pronunc.: *Inton.:*		
Positive elements:		*Positive elements:*		
2. Communicative output (quant. + qual.)	average score: /5	5. Vocabulary		score: /10
Task 1:		*A. Words (general purposes)* *B. Word combinations (valency, collocat., express.)* *C. Richness, precision y variation* *D. Communicative functions*		
Task 2:				
Task 3:				
Attitude / Non-verbal commun.:				
Positive elements:		*Positive elements:*		
3. Comprehension	score: /5	6. Grammar + pragmatic-discursive structures		score: /10
		A. Morphology *B. Syntax* *C. Pragm.-disc. struct. (register, struct., coh., conjunct., compensatory strateg. and interact.)* *D. Variation*		
Positive elements:		*Positive elements:*		

COMPENSATORY ELEMENTS (portf., progress, attitude; +1 > +3) + | PENALIZING ELEMENTS (portf., progress, attitude; -1 > -3) -

Table 3. Template for oral competences in a B2 level course in a translation and interpreting curriculum. Second "evaluation layer".



Template Oral evaluation AL B2 Name: Evaluator: Total: /40

A. "TASK"	total "A": /15	B. "TEXT"	Global AAA B C D E F	total "B": /25
1. Fluency and quantity	score: /5	4. Pronunciation + intonation		score: /5
5= object. totally reached. / 4= ... largely (< 2 vacillat.) / 3= ... partially. (2-4) / 2= ... 50-50 (communic. disrupted) / 1= ... not at all ... / 0= incomprehensible		*Pronunc.* 3= 0 errors / 2= 1-3 errors / 1= 4-6 errors / 0= >6 errors		
Comments:		*Inton.* 2= 0 errors / 1= 1-3 errors / 0= >3 errors		
		Comments (underline errors level A):		
Positive elements:		*Positive elements:*		
2. Communicative output (quant. + qual.)	average score: /5	5. Vocabulary		score: /10
5= task performed without errors / 4= in a relevant & coherent way (<2 errors) / 3= partially... (2-3) / 2= 50-50 (4-6) / 1= insufficiently. (>6) / 0= incomprehensibly		*A. Words (general purposes):* 3= no errors / 2= 1-3 / 1= 4-6 / 0= >6		
Task 1:		*B. Word combinations (valency, collocat., express.):* 3= no errors / 2= 1-3 / 1= 4-6 / 0= >6		
Task 2:		*C. Richness, precision y variation:* 3= very rich; 2= relatively...; 1= not enough...; 0= only basic words		
Task 3:		*D. Communicative functions:* 1= correct; 0=incorrect		
Attitude / Non-verbal commun.:		*Comments* (underline errors level A):		
Positive elements:		*Positive elements:*		
3. Comprehension	score: /5	6. Grammar + pragmatic-discursive structures		score: /10
5= total / 4= quasi total (<2 probl.) / 3= partial (2-3) / 2= limited (4-6) / 1= insufficient (>6) / 0		*A. Morphology:* 3= no errors / 2= 1-3 / 1= 4-6 / 0= >6		
Comments lex. (concr./abstr.) gram. / pragm.-disc. (underline errors level A:		*B. Syntax:* 3= no errors / 2= 1-3 / 1= 4-6 / 0= >6		
		C. Pragm.-disc. struct. (register, struct., coh., conjunct., compensatory strateg. and interact.): 3= very rich and correct; 2= relatively...; 1= not enough...; 0= not present		
		D. Variation: 1= sufficient; 0=insufficient		
		Comments (underline errors level A):		
Positive elements:		*Positive elements:*		

COMPENSATORY ELEMENTS (portf., progress, attitude; +1 > +3) + | PENALIZING ELEMENTS (portf., progress, attitude; -1 > -3) -

Table 4. Template for oral competences in a B2 level course in a translation and interpreting curriculum. Third "evaluation layer".

References

Buyse, K. (2012). Criterios y plantillas para la evaluación del español para fines específicos. El español de las profesiones. Artículos seleccionados del IV Congreso Internacional de Español para Fines Específicos. Congreso Internacional de Español para Fines Específicos (pp. 186-200). Amsterdam, 18-19 November 2011. Madrid: Ministerio de Educación, Cultura y Deporte.

Buyse, K. (2013). "Come and sit here next to me". Towards a communicative assessment of oral language skills (submitted for publication in Language Learning in Higher Education, 2).

Cabré, M.T. & Gómez de Enterría, J. (2006). La enseñanza de los lenguajes de especialidad. La simulación global. Madrid: Gredos.

CITO/SLO. (2010). Toetsen en beoordelen met het ERK. Arnhem/Enschede.

Dochy, F. & Gijbels, D. (2010). Evaluatie. In S. Janssens (Ed.), Leren en onderwijzen (pp. 121-175). Leuven: Acco.

Gipps, C.V. (1994). Beyond testing: Towards a theory of educational assessment. London: The Falmer Press.

Gijselaers, W. (2007). Talking about My Generation. Keynote at EARLI 07: European Practice-based and Practitioner Research conference on Learning and Instruction for the new generation, Maastricht, 14-16 November.

Keeves, J.P. (1994). Methods of assessment in schools. In T. Husén & T.N. Postlewhaite (Ed.), International Encyclopedia of Education (pp. 362-370). Oxford & New York: Pergamon Press.

Luoma, S. (2004). Assessing speaking. Cambridge: Cambridge University Press.

O'Sullivan, B. (Ed.). (2011). Language testing: theories and practices (Palgrave advances in linguistics). Basingstoke, UK: Palgrave Macmillan.

Parrondo Rodríguez, J.R. (2004). Modelos, tipos y escalas de evaluación. In J. Sánchez Lobato & I. Santos Gargallo (Ed.), Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2)/lengua extranjera (LE) (pp. 967-982). Madrid: SGEL.

Robles Ávila, S. & Cárdenas Bernal, F. (2006). La enseñanza del español como lengua extranjera a la luz del marco común europeo de referencia. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Sluismans, D. & Dochy, F. (1998). Alternatieve toetsmethoden in studentgericht onderwijs. Tijdschrift voor Hoger Onderwijs, 16(4), 229-233.

Van Iseghem, J. (2010). Didactiek mondelinge vaardigheden. Leuven: KU Leuven.

Cristiana Cervini, Monica Masperi, Marie Jouannaud & Francesca Scanu

Université Stendhal, Grenoble, France
Università di Bologna, Bologna, Italy

cristiana.cervini@unibo.it

Defining, Modeling and Piloting SELF, a New Formative Assessment Test for Foreign Languages

Bio data

Cristiana Cervini, PHD in educational linguistics, teaches 'applied linguistics to foreign language learning' at the LILEC Dep. (University of Bologna). Her present studies focus on assessment and evaluation and on CALL systems for hybrid and self-learning. She is currently in charge of the SELF research and development actions in the frame of the INNOVA-Langues project (LANSAD, Grenoble 3).

Monica Masperi is a Senior Lecturer in Linguistics and Didactics at Stendhal University, Grenoble, France. At the head of the LANSAD department (Languages for non-specialists) since 2004, she is also the scientific director of the INNOVA-Langues project. Her current research as a member of the LIDILEM research lab focuses on Italian didactics, plurilingualism and the use of technology in language teaching and learning.

Marie-Pierre Jouannaud has been a foreign language instructor for more than twenty years and currently teaches English linguistics and teaching methods at Stendhal University, Grenoble, France. She has a Masters degree in linguistics and is about to start a PhD in applied linguistics. Her areas of interest include acquisition of grammar, blended learning and the development of learner autonomy. She is the English item writer coordinator in SELF.

Francesca Scanu is a teaching assistant in the LANSAD department of Stendhal University - Grenoble 3. She has a Master's degree in Didactics of Italian as a Foreign Language (University for Foreigners, Siena) and is actively engaged in the creation of language learning paths. Within the INNOVA-Langues project, Francesca is currently working on the creation of SELF.

Short paper

Assessing foreign languages in higher education: state of the art and preliminary results

Designing a new foreign language test requires defining what we mean by language and language use (Bachman, 1990). The construct that we use is the cornerstone guiding us when we create original test items and when we design the general architecture of the test. What are the characteristics of the intended test-takers? What do we take 'communicative competence' to be? How can we translate all of this into test items that will be administered online? The Common European Framework of Reference for languages (henceforward CEFR) can guide us but cannot tell us how to anchor our items to the descriptors of each of the skill levels while staying true to task-based and action-oriented approaches to language teaching, and to the role of the learner as a social actor (ALTE, 2011; Weir, 2004).

The purpose of this talk is to describe two aspects of this work in progress:
i) how our team's linguistic, discursive and contextual choices were guided by the principle of situational and interactional authenticity on the one hand, and by the attempt to integrate competences on the other; ii) what the preliminary stages of the piloting process tell us about our students' socio-biographical characteristics and the validity of our test.
SELF (Système d'Evaluation en Langues à visée formative) will eventually cover three language skills (listening, reading and writing[1]), but the first stage of its development focused on listening, because of the high correlation we observed between oral comprehension level and success in foreign language tests. Developing a test to be used in institutional settings implies a series of inevitable constraints due, on the one hand, to the possible wash-back effects on learning and on teaching models and, on the other, to the high number of test-takers taking the test at the same time.

How can we overcome the limits of computer-assisted testing and standardization in foreign language evaluation?
Communicative and task-based/action-oriented approaches require taking into account pragmatic and even sociolinguistic variables, whereas standardized automatic scoring seems more compatible with the testing of discrete linguistic knowledge associated with more traditional methods (morphosyntax, spelling, phonology and lexis).

If we wanted to align closely with the CEFR, we would need to include efficacy and communicative relevance in our analyses, both for monologic (spoken production) and dialogic texts involving two or more speakers (spoken interaction), especially at the higher levels. It is not easy, however, to integrate these fundamental aspects within the constraints of computer-assisted testing: not only will the test have to be automatically corrected (it will be administered to hundreds of students more or less simultaneously during registration week), but it will also need to be relatively short.
Sociolinguistic competences are another source of difficulty. According to the CEFR, being able to identify regional dialects as well as elements of a country's popular culture is one of the skills displayed at higher levels. Unfortunately, considerations of equity prevent us from using dialects and any questions vulnerable to interpretations of cultural stereotyping that might offend some learners or put them at a disadvantage without having anything to do with linguistic competence (Kunnan, 2010).
Some compensatory measures are obviously needed to make sure that our construct of linguistic competence is compatible with the constraints of online testing.

Corpus-based, authentically grounded and home-made items
We define items as 'minimal units of content allowing verification of a linguistic objective'. Most items are self-contained, but their identity is also defined in contrast with or in relation to other items within the system or sub-system they belong to. Of course, the audio document that each item uses determines to a great extent the characteristics of the item. In our case, the three main sources used are:

- Home-made: the item writers create a text (dialogue, news item, …) centered around a communicative, lexical or morphosyntactic element that can thus be specifically targeted because it is hypothesized to be a critical component of linguistic competence at a given level;
- Corpus-based items whose associated audio text came from transcribed corpora of oral language (e.g. the LIP corpus: « Lessico di frequenza dell'italiano parlato »);
- Finally, items whose audio comes from an authentic document whose original purpose had nothing to do with teaching or testing (for example public announcements in a train station).

[1] It might be useful to indicate that exercises should be self-corrective hence we would focus on "limited production" (i.e.: short answers, discourse completion task, etc.).

In order to focus more closely on communicative competence in interaction , one of our item types for listening involves the test taker having to choose the best and most appropriate response for the next turn in an on-going conversation (these items are in a way similar to Dialang's register/appropriacy items in their writing construct and the DCT (discourse completion task) type of exercise).

Situational and interactional authenticity, integration of competences

Language in use does not separate competences, but discrete point testing does. We have tried as much as possible to balance these two contradictory requirements by designing an "identity card" to help us describe, create and classify items. This tool is also essential for the training of new item writers as well as for research and evaluation.
Here are the definitions we use for authenticity and integration of competences:
"Situational authenticity" refers to the accuracy with which tasks and items represent language activities from real life. "Interactional authenticity" refers to the naturalness of the interaction between test taker and task and the mental process which accompany it. [...] To make an item or task more situationally authentic, the key feature of the real life task must be identified and replicated as far as possible".
"Integrating competences": when we are designing a test task, it is important to be clear about the balance between competences needed for a successful response. Some competences will be more important than others - this will form the focus of the task" (Manual, 2011).
At the micro level, enriching each item with exhaustive contextual details aims to make up for the loss of paralinguistic information that is typical of naturally occurring exchanges. More specifically, each item is composed of a series of elements including "contextual clues" given to the test-taker (usually the place where the scene takes place or a short introduction to the topic), the pedagogical direction (explaining what the test taker has to do), which is separate from the functional direction or prompt (the technical operation associated with the item type, such as "choose the correct answer" for a multiple choice item). These indicators complete the input (the audio clip) and the answers (the key and the distractors).

A modular structure that is flexible enough to adapt to different uses

One of the most interesting aspects of the projected SELF system is the variety of purposes for which it is intended. It is designed to be used with non-specialist students (i.e. students who are not majoring in languages) who are taking FL courses to fulfill their foreign language requirement or as an elective. The delivery models for these courses range from face-to-face learning with enriched online content to blended learning or hybrid courses and fully online tutored courses.
SELF will be used as a placement test ("a test administered in order to place students in a group or class at a level appropriate to their degree of knowledge and ability" (ALTE Multilingual Glossary, 1998)) as well as a diagnostic test ("A test which is used for the purpose of discovering a learner's specific strengths or weakness. The result may be used in making decisions on future training, learning or teaching" (ALTE Multilingual Glossary, 1998)). This kind of formative assessment helps students improve their capacity for self-assessment, which, along with greater awareness of their weak and strong points, is the first step towards autonomy. It will also provide information to the tutors who will guide the students during their personalized self-directed online training sessions.
SELF is thus conceived as a modular structure focused on the assessment of three language activities: listening, easing and writing. This modularity will allow for flexible uses. As a placement test, its administration should not exceed 50 minutes or so. As a diagnostic test, it will be possible to assess each skill separately. In this case, its main purpose will be formative, designed to foster learners' autonomy.

What is listening comprehension? items and the cognitive operations they imply

Defining the construct of listening means that we have had to reflect on the stages of the process, the subskills and the cognitive operations involved. Different foci are possible: 'listen and perceive', for bottom-up phonetic and prosodic processing, 'listen and

understand', i.e. (re)constructing and interpreting meaning to grasp the message and 'listen and interact', because efficient and relevant interaction must be based on adequate reception and comprehension of the other speaker's discourse (Cornaire 1998, Nobili 2006).
As far as the difficulty of the task (listening comprehension theory) and the difficulty of items (testing theory) are concerned, the descriptors of the CEFR lead us to identify several contributing factors, namely:

1. the linguistic characteristics of the audio input, i.e. speech rate, pauses, hesitations (phonetic characteristics), or fragments, atypical word order, variation typical of spontaneous oral discourse (morphosyntax), information density linked to the type of text, length and narrative organization of the text (discursive characteristics);
2. the intrinsic difficulty of the exercise type (MCQ, close, T/F, matching, reordering, …), time available to complete the task, use of tools such as the possibility of note-taking;
3. the personal characteristics of the test-taker: ability to make predictions, activation of background knowledge, capacity for sustained attention, verbal working memory, and attitude;
4. the cognitive processes involved, such as listening for gist, listening for details, inferencing about the context (where the scene takes place, who the speakers are, …), recognizing communicative intent (and its effects), identifying mood, register, etc…

The "ID card" of each item, a tool we have perfected as our research progressed, has helped us to:

- Make explicit the focus of the item, so that no ambiguity remains as to what is being tested;
- Raise item-writers' awareness of item characteristics and simplify quality control;
- Facilitate access to specific items with the search engine (by their focus, length, text type, domain, …) and trace their behavior after piloting.
- Make sure there is a variety of focus (lexical, morphosyntactic, communicative), text types and length (long or short, authentic, modified or invented, monologic or dialogic, announcements, instructions, conversations, …), speech rate (fast, medium or slow), language variety (standard or non standard), domains (public, private, professional and educational) and that the test-taker is cast in different roles (listener/eavesdropper or participant).

Two types of tools have turned out to be very useful for the item writers and revisers: reference level descriptions (e.g. 'Il Profilo della Lingua Italiana' for conceiving items in Italian) and oral and written language corpora (such as LIP, CORIS/CODIS, LABLITA).

A tiered approach to the validation of an adaptive test

According to Doucet, the validation process can be likened to an accumulation of converging data until we are convinced that the approach we have chosen is well founded: « Lors du processus de validation, on parlera d'accumulation convergente de données jusqu'à ce que l'on soit convaincu du bien-fondé de l'approche choisie » (Doucet, 2001)
Validatity and reliability are two concepts that come from psychology, where they have been in long use. Reliability implies the stability and coherence of test results over separate administrations (« selon le principe de fiabilité, le résultat d'un test doit rester le même entre deux occasions d'évaluation rapprochées », Lussier & Turner, 1995), whereas validity measures the convergence between the aim of the test and its contents (« le principe de validité sert à vérifier si le test mesure effectivement les performances qu'il cherche à mesurer », ibid.).
We have imagined several steps in the revision and validation process, both qualitative and quantitative, starting with the A2 level listening comprehension items. The qualitative analysis starts with a revision of the items by an expert who is not involved in the writing process, followed by 'think aloud protocols' (TAP) with a few learners selected

because their level in the skill tested corresponds to that of the items studied (this will also enable us to check that the test has face validity for learners).
Scientific validation is then completed by a guided and controlled pilot study where we administer the test to a small group of students of similar ability who serve as a preliminary control group. This is followed by a pretesting phase, in which we observe and interpret statistical data obtained from a much larger pool of learners.

For the think aloud protocol, observing results from two very different groups, one in a second language context (immersion), the other in a foreign language context (non immersion), will allow us to compare their reactions. The first group are volunteer Italian as a foreign language (Lansad) students in Grenoble, France, pretested A2 with Dialang, and the second are Erasmus students in Bologna taking Italian language courses in the Language Center, also pretested A2 using an in-house test. This protocol will help the test designers verify that the thinking processes used correspond to the hypothesized construct, and observe the behavior of the items as well as the impact of our linguistic choices on participants with very heterogeneous biographical and educational backgrounds.
We have focused on the following questions:

- Perceived difficulty of the item (high/ medium/ low), which can be compared with the answer given, whether right or wrong;
- What did the learner find difficult: comprehension of the audio document, of the stem, the answers (key and distractors), the prompt, or was there interference with the native language(s)?
- Is the speech rate of the document too fast, the lexis or structures too complex, the details too hard to follow?
- What is the student's level of confidence in their answer (25/ 50/ 75/ 100%), and would the student have chosen to answer if incorrect answers deducted points out of the total score?
- Comments on the type of protocol and received instructions.

The first results collected during the qualitative analysis have uncovered a bias in one specific multiple-choice item. The oral input consisted in a dialogue between a mother and her son and the question was focused on the interpretation of the mother's feelings. The TAP analysis showed that the mother's behaviour (anxious, angry or thoughtful?) was subject to diverging personal and cultural interpretations, causing a bias. This evidence has led the item-writers to proceed with a second revision and modification of the exercise.

Conclusions: research modules and their perspectives
The SELF evaluation system, a multidimensional tool, will evolve along predetermined lines, the next step being the development of the following modules (or 'research bricks'):

- Self-evaluation module: this step precedes the test proper and has a functional as well as a formative purpose. Functionally, this will allow us to start the test with items closer to the level of the learner, and reduce the time necessary and the stress or boredom associated with excessively hard or easy questions for the test-takers. Formatively, it would be interesting to link the self-evaluation results to the final diagnosis. We are exploring different modalities for self-evaluation, including metacognitive can-do statements ("Je suis capable de…"), benchmarked samples to which the learners can compare their performance, or using ideas developed in portfolio assessment projects, which might mitigate the drawbacks of questionnaires and can-do statements.
- Formative feedback module: feedback is a central issue for a formative test, both for the student and the instructor. This implies deciding what information the student needs to see after finishing the test, and what needs to be stored for the long term, or perhaps permanently, perhaps in a "personal profile" page.

- Scoring module: The protocols we choose for our exercises, the competence levels aimed at, and the use of dichotomous vs. polytomous items will all impact scoring procedures.

References

Bachman L. F. & Palmer A. S. (1996). Language testing in practice: Designing and developing useful language tests, Oxford: Oxford University Press.

Bachman L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press .

Cornaire C. (1998). La compréhension orale, CLE International: Paris.

De Mauro T., Mancini F., Vedovelli M. & Voghera M. (1993). Lessico di frequenza dell'italiano parlato, Etas Libri: Milano.

Doucet P. (2001). Pour un test utile. ASp, 34. URL : http://asp.revues.org/1696; DOI : 10.4000/asp.1696.

Kunnan A. (2010). Statistical Analysis for Test Fairness. Revue française de Linguistique appliquée, 16.

Lussier D. & Turner C. E. (1995). Le point sur l'évaluation en didactique des langues, CEC - Centre éducatif et culturel : Montréal.

Manuel pour relier les examens de langues au Cadre européen commun de référence pour les langues. (CECR). (2009). URL http://www.coe.int/t/dg4/linguistic/manuel1_fr.asp.

Manual for Language Test Development and Examining (for use with the CEFR), produced by ALTE (Language Policy). (2011). URL : http://www.coe.int/t/dg4/Linguistic/ManualtLangageTest-Alte2011_EN.pdf.

Nobili P. (2006). Oltre il Libro di Testo. Carocci : Roma.

Rubin J. (1994). A Review of Second Language Listening Comprehension research. The Modern Language Journal. 78, ii.

Weir C. J. (2004). Limitations of the Common European Framework of Reference. Developing Comparable Examinations and Tests, URL: http://ltj.sagepub.com/content/22/3/281.abstract.

Yu-Hua Chen, Shaida Mohammadi & Veronica Benigno

Pearson, London, United Kingdom

yu-hua.chen@pearson.com - shaida.mohammadi@pearson.com - veronica.benigno@pearson.com

What and How Many Words Do We Need? Critical Considerations when Developing a CEFR Vocabulary List: Size, Depth, and Growth

Bio data

Yu-Hua Chen trained as a test developer and corpus linguist. Her research interests focus on how corpus analysis can facilitate or validate the approaches to teaching and testing language skills. Her doctoral dissertation was selected as a finalist for Jacqueline Ross TOEFL Dissertation Award 2012. Publications include Lexical Bundles in L1 and L2 Academic Writing (2010, ao-authored with Paul Baker) and The Academic Collocation List (2013, co-authored with Kirsten Ackermann).

Shaida Mohammadi is currently working as a Test Development Manager with Pearson. Her research interests include assessment, research methods, materials development and teacher education.

Veronica Benigno is currently employed by Pearson UK in the position of Research Manager. She manages external organizations' research projects on PTE Academic (Pearson Test of English Academic) and conducts internal research on language assessment issues. Doctor Europaeus in Linguistics and Didactics, she is an active researcher in second language acquisition, corpus linguistics and vocabulary studies. She has worked as teacher of Italian at the University of Palermo and as lexicographer, contributing to the publication of a combinatory dictionary of Italian.

Abstract

Since the launch of the CEFR, there have been various attempts to develop a vocabulary list aligned to it. However, little effort has been made to translate the theoretical frameworks or empirical findings of vocabulary research into the development of such listings. For example, recent studies have concluded that knowledge of 8,000-9,000 word families is necessary for reading authentic English materials and perhaps 5,000-7,000 families for oral communication (Nation, 2006; Schmitt, 2008). In comparison, most of the available pedagogical vocabulary listings contain a much smaller repertoire. The 'multi-facetness' of vocabulary knowledge (Read, 2000) should also be acknowledged when developing a CEFR-aligned syllabus. There is a whole gamut of 'knowing' a word, from recognizing only one context-dependent sense to frequently using it with a variety of meanings. To address the gap between vocabulary research and alignment with the CEFR, this paper critically reviews the construct of vocabulary as a language ability and focuses the discussion on three main dimensions, i.e. vocabulary size, depth and growth.

To establish the relationship between vocabulary knowledge and the CEFR, we will demonstrate an innovative approach using information from language testing statistics and L1 corpus frequency, taking into account the above three dimensions. The rationale is that L1 corpora offer a comprehensive list of words from authentic text produced by native speakers, while L2 learners' performance in a language test, specifically their responses to different item types that assess various facets of vocabulary, can be used to gauge the extent to which they know the words in different contexts. Preliminary findings

from a pilot study will be reported, and the results confirm the complexity and multi-facetness of vocabulary as a construct. The pro and cons of using various sources to inform vocabulary syllabus design and the implications for CEFR alignment will also be discussed.

Short paper

Introduction

The Common European Framework of Reference (CEFR) itself contains a repository of can-do statements for various skills and aspects of language use, rather than explicit language knowledge of required lexis and grammar. In this paper, we will not debate whether a traditional vocabulary list should be developed on the basis of CEFR because different projects like this have been undertaken around the world. Instead, the focus of the paper will be on how such a link can be established, drawing on relevant vocabulary research and available vocabulary listings.

A vocabulary list can be a simple list of words only or can comprise various components, including headwords and any other relevant information such as parts of speech, meaning, usage, collocations or phrases, etc. under the same headword. For a pedagogical syllabus to be meaningful, however, learner level is arguably the most important piece of information that needs to be taken into account because it is the backbone that underpins the range of language that learners are expected to be able to function with at that level. Unlike dictionary compilation, where exhaustive information about a word can all be listed under one entry with some sort of ordering principles, the information presented in a pedagogical vocabulary list needs to be selective – only what is considered appropriate to that level should be included. And it is this level of appropriacy that makes developing a vocabulary list aligned to the CEFR such a mammoth undertaking.

One of the major issues in aligning vocabulary with can-do statements in the CEFR is that language functions can usually be fulfilled through more than one expression, thus opening up a broad range of possible lexical items. Therefore one implication of developing a vocabulary list on the basis of the CEFR is that such a direct linkage between functions and lexis may not be easy to establish, and may be even more difficult when one attempts to turn the required lexis into something comprehensive enough to be a pedagogical list. Instead of compiling such a list, the British Council and EQUAL [1] (2011) provide a core inventory, including exponents of functions, notions, grammar and lexis, and the exponent of lexis is illustrated in contexts under topic categories. An example of A1 lexis is shown below:

Personal information

- She's married and has three children.
- I am 26 years old, single and I work in a bank.
- He's an engineer.

On the other hand, the English Profile project (Capel 2010, 2013) tackles the issue by extracting learner evidence at predetermined CEFR levels from a large learner corpus and has them reviewed by experts in the field and/or compared with L1 corpus data. A comprehensive list of vocabulary has been developed and each entry is exemplified with both L1 use and L2 evidence of written samples under testing conditions. However, careful scrutiny of its documentation suggests that the total vocabulary size represented in the final listing is likely to have underestimated the demands of the lexis required to perform the functions specified for the highest level, C2. In addition, relying on learners' written samples as the main source of evidence might also have biased the listing

[1] EQUAL stands for the European Association for Quality Language Services.

towards the more productive side of lexical ability, as opposed to a fair representation of receptive and productive lexicon.[2] Both vocabulary size and the 'multi-facetness' of vocabulary knowledge (Read, 2000), i.e. depth, will be discussed in greater detail in the remainder of this paper.

In addition to vocabulary size and depth, this article will also address another key dimension of lexical ability, i.e. growth, which refers to the determination of the cut-off points between CEFR levels. An overview of various approaches and sources of evidence that such a vocabulary listing can draw on will be provided. As part of a broader proposed framework of syllabus development, some findings from a pilot study using test statistics and corpus frequency will be described.

Vocabulary size

Over the decades, researchers have been striving to answer this question: How many words do learners need to know? Well-educated native speakers are estimated to know approximately 20,000 word families (Goulden, Nation & Read, 1990; Zechmeister, Chronis, Cull, D'Anna, & Healy, 1995; Nation and Waring, 1997)[3], while the Longman Dictionary of Contemporary English includes 230,000 words, phrases and meanings and the Longman Phrase Bank comprises 220,000 entries for word combinations (Pearson Education, 2009). For second language learning, Nation, in his seminal paper (2006), analysed authentic texts, including novels, newspapers, radio programmes and interviews, and concluded that a lexical threshold of 8,000 to 9,000 word families is recommended for reasonable comprehension of authentic written text, and a vocabulary of 6,000 to 7,000 for spoken text. These lexical thresholds were subsequently supported by Schmitt (2008), Laufer and Ravenhorst-Kalovski (2010) and Schmitt and Schmitt (2012), while Van Zeeland and Schmitt (forthcoming) contend that 2,000-3,000 word families should suffice for adequate listening comprehension, though only informal narrative passages were used in their experiments. Although it is reasonable to assume that the estimated figures above approximate to the vocabulary size of very proficient L2 speakers, very few studies in vocabulary research use the CEFR as a reference for learner proficiency.

If we turn to the level descriptors in the CEFR and focus on the highest level of C2, as shown below, it seems reasonable to assume that the wording for C2, such as 'a very wide range', 'very broad lexical repertoire', 'finer shades of meaning' and 'to differentiate and eliminate ambiguity', suggests that the vocabulary size of the most proficient L2 speakers at C2 should at least reach the highest estimated figures in the above vocabulary research, if not more.

RELEVANT QUALITATIVE FACTORS FOR RECEPTION C2 (COE, 2009, p.143)
Can understand a very wide range of language precisely, appreciating emphasis and differentiation. No signs of comprehension problems.

Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.

RELEVANT QUALITATIVE FACTORS FOR PRODUCTION C2 (COE, 2009, p.149)
Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.

[2] According to Capel (2013), the English Vocabulary Profile does not acknowledge a distinction between receptive and productive vocabulary until the C1 level.
[3] As will be discussed later, various definitions and methodologies make it difficult to compare the vocabulary size across different studies, but the figure of 20,000 word families reported here are believed to be a more realistic estimate (see Nation, 1993, 2006).

Interestingly, Milton and Hopkins (2006) estimate that 4,500-5,000 word families is enough to reach CEFR C2 and the highest level of Cambridge English Test CPE (Certificate of Proficiency in English). The English Vocabulary Profile is, to the best of our knowledge, the only large-scale project that has been completed in an alignment exercise for all six levels of the CEFR, and the final listing also reports a much lower number of entries – just under 7,000 headwords and approximately 15,000 senses and phrases in total (Capel, 2013).

Note that the operational units for counting purposes, i.e. words, word families or headwords, vary in different studies and projects, and this has a major impact on estimation of the required vocabulary size. The notion 'word family', which refers to a basic or root form, its inflections and derivatives, appears to be more widely acknowledged as a psycholinguistically valid unit for potential learning purposes (see Bauer and Nation, 1993, for a comprehensive 7-level taxonomy of word family categorization). The rationale is that if learners know about the word friend, for example, it is most likely that they will also know more or less about, or be able to guess, the meanings of friends, befriend, friendly, friendliness, although the taxonomy itself does not necessarily represent a linear relationship to the degree to which these associated forms can be easily recognizable for learners (see a critical discussion of taxonomy in Gardner 2007).

Despite the confusion arising from terminology and operational units, there is no doubt that a significant gap exists between the lexical threshold generally acknowledged in vocabulary research and the actual alignment of vocabulary size to the CEFR. The lexical thresholds recommended by Nation embrace the notion of word family, which is actually an operational unit covering a much broader range of word members than headwords (supposedly comprising bare and inflective forms only) used in the English Vocabulary Profile. The gap in vocabulary size difference, therefore, is probably much greater than it appears on the surface.

Vocabulary Depth

Vocabulary depth refers to the multi-faceted aspect of vocabulary knowledge, as there is a wide range of views on the definition of 'knowing' a word, from recognizing only one context-dependent sense to frequently producing it with a variety of meanings (Read, 2000). There is general agreement that the interpretation of vocabulary depth is diffuse and difficult to determine which aspects of vocabulary depth are more important than others and should be emphasized (Read, 2004, 2007; Milton, 2009), which highlights the need for further research on what dimensions should be considered when defining depth. In its broader context, vocabulary depth may entail receptive and productive knowledge (aka active and passive vocabulary, see: Laufer and Baribakht, 1998; Laufer, 1998) as well as knowledge about the associated behaviours of words, e.g. polysemy, homonymy, multiword expressions including collocations, phrases, idioms and any other formulaic expressions. It is suggested that there is little merit in eliciting all aspects of a set of words and that a better option is the testing or examining of only some selected aspects of word knowledge (Schmitt, 1998). This has resulted in a focus on receptive and productive aspects of depth.

Receptive vocabulary is known to be typical larger than productive vocabulary (Schmitt, 2008), although some argue that learners' vocabulary depth knowledge increases with their proficiency development and thus knowledge of vocabulary size and depth might converge at higher levels (Vermeer, 2001; Akbrian, 2010). Yet, at least for lower levels, this suggests that productive knowledge is built on receptive knowledge, which explains the slower rate at which productive vocabulary develops compared to receptive vocabulary (Zhou, 2010).

For another aspect of vocabulary depth, polysemy and homonymy, the English Vocabulary Profile has done a fantastic and meticulous job by considering the 'sense'

level of individual words and multi-word expressions. However, one issue with using learner evidence to inform the development of a vocabulary syllabus is that learners tend to avoid using some types of multi-word expressions, such as phrasal verbs (Dagut and Laufer, 1995; Liao and Fukuya, 2004), whereas learners also tend to overuse other types of multi-word expressions, such as discourse markers, e.g. on the other hand, at the same time, which diverts from the norm found in L1 language (Chen and Baker 2010). Learner evidence, therefore, is probably not the best source in this regard as learners do not seem to produce multiword forms in the way that they are expected to, but this does not mean that they should not be taught these multiword expressions earlier.

It is very important to include multi-word units in the syllabus because, nowadays, it is broadly acknowledged that words are not learned as single lexical units out of context. The nature of language is phraseological (Hoey, 2005): speakers make use of prefabricated pieces of language, chunks, multi-word expressions, collocations, pragmatic units. Take collocations for example. Studies of the relationship between collocational use and proficiency levels (Boers et al., 2002; Bonk, 2011; Gitsaki, 1999) suggest that collocations are better mastered as proficiency increases. Collocational use is therefore an index of lexical depth, a dimension which is often neglected in studies of vocabulary, as mentioned above. Other studies (Durrant and Schmitt, 2009; Ellis, 2001; Laufer and Waldman, 2011) suggest that collocational knowledge clearly distinguishes native speakers from learners. Although there may be a positive correlation between frequency and knowledge of collocations, it seems that it is not just a matter of frequency: native speakers would regard very fixed collocations as communicatively essential, even if they occur with very low frequency (Benigno, 2012). The debate on learning and teaching collocations is ongoing, and of primary importance to the understanding of how language works. Yet, when compiling a vocabulary list, there is no doubt that multi-word expressions, such as collocations and phrases, are essential, and perhaps L1 corpora would be a better source for identifying the multi-word lexical items that learners should know or learn rather than using L2 evidence.

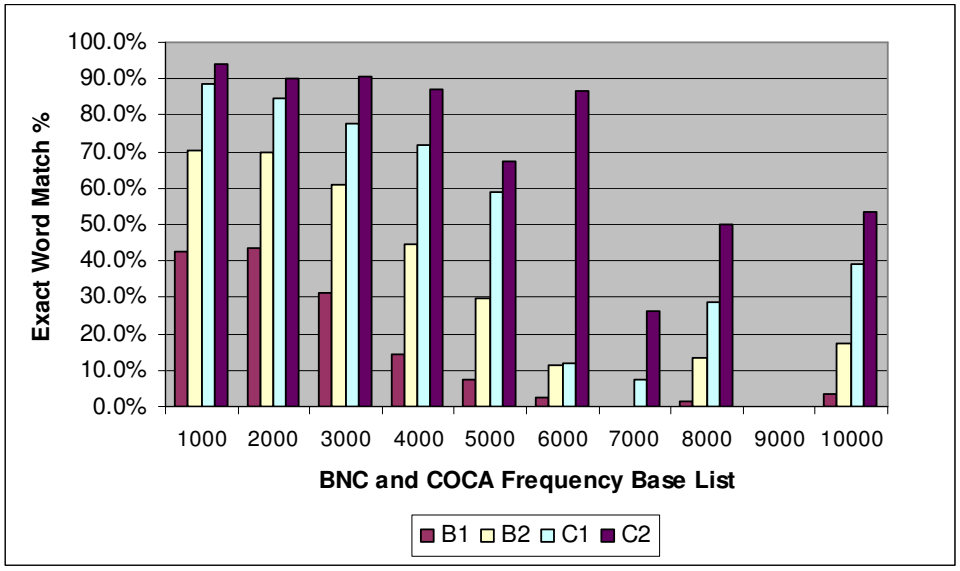
Vocabulary Growth

Vocabulary growth refers to the cut-off point of language development, i.e. CEFR levels in the current discussion. For productive vocabulary, it is possible to make a judgement relying on evidence from learners' spoken or written samples at a given level, and this is the approach adopted by the English Vocabulary Profile, A1 to B2. For receptive vocabulary, test takers' performance in the Vocabulary Levels Test (Nation, 2001; Schmitt, Schmitt, and Clapham, 2001) or the Vocabulary Size Test (Nation and Belgar, 2007) may be used to inform syllabus development if typical learners at specific CEFR levels can be identified.

As empirical validation is equally important as a theoretical framework for developing a vocabulary syllabus, we are currently working on an innovative approach involving combining multiple sources to better inform syllabus development and we will present some preliminary findings from a pilot study here. This pilot study is part of a broader project, where test statistics and L1 corpora are used to shed some light on the construct of vocabulary knowledge. In one specific task, dictation, test-takers hear a sentence and write it down exactly as they hear it. This task assesses the ability to comprehend the text with aural input and produce the corresponding written output, hence a measure of assessing integrated vocabulary ability. For the pilot study, approximately 25,000 test takers' responses to 26 items were collected, and a list of all the written words produced by test-takers at different CEFR levels[4] was generated and compared against the frequency base lists of word families developed by Nation (2012). These word family lists, divided by frequency rankings per 1,000 families, were compiled using the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), which represent native-speaker British English and American English in both written and spoken

[4] CEFR levels were determined by test takers' overall scores in PTE Academic, an academic English test.

registers. The taxonomy of word families follows the system established by Bauer and Nation (1993). The results of mapping test-takers' vocabulary for the dictation task with BNC/COCA frequency lists are presented below. Note that test-takers' vocabulary ability is presented as a percentage of correct words produced, as only those words with correct orthographic form were counted. As can be seen, a nearly perfect linear relationship between vocabulary size and CEFR level emerges from the graph, which suggests that C2 learners have a lexicon of up to 10,000 word families, whereas B1 learners start to struggle after the first 3,000 word families. More evidence will be provided at the time of the conference.



Despite its great potential, there are still issues with this approach. The test items contain a lot more words from frequency lists for the first 2,000 families and very few from low frequency bands. For example, no word from the 9,000th word family was covered, and there is thus no evidence from that very specific range. In addition, multi-word expressions, unfortunately, cannot be accounted for when using automated corpus tools like Range. This is possible, however, with manual extraction if a list of multi-word expressions being tested in the task can be compiled.

As this is just a pilot study from an ongoing project, more task types that tap into different dimensions of lexical ability will continue to be investigated. These tasks include two more integrated types, Listen and Highlight Different Words, Listen and Type Missing Words, and the more traditional Essay Writing. It is hoped that they will provide further insights into our understanding of vocabulary size and the 'multi-facetness' aspect of vocabulary ability in relation to CEFR levels. In the long term, individual lexical items from L1 corpus frequency lists will also be annotated with statistics from live language tests covering both receptive and productive skills to determine to what extent L2 learners know a word at a specific level. We are also continuing to consult ELT materials and experts to identify the lexical gap that might occur between L1 and L2, especially at lower levels, and to include, for example, lexical items that are often derived from 'tourist' or 'survival' English but have lower frequency in L1 corpora (e.g. 'beef', 'milk').

Concluding remarks

In this article, we argue that various dimensions of vocabulary as a construct should be acknowledged in an exercise aligning a vocabulary syllabus to the CEFR. We also evaluate the pros and cons of various sources which can be used to inform the development of such a vocabulary syllabus. As no single source can accommodate all the complexity of defining vocabulary ability, we have demonstrated a promising innovative

approach which combines L1 frequency lists and test statistics from items that tap into the 'multi-facetness' of lexical ability in order to inform the vocabulary size and depth of knowledge of learners at individual CEFR levels. It should be noted that this is not a circular practice – simply recording what learners know and feeding it back to the syllabus – as one might suspect. As Chen (2011) points out, the purpose and audience should be defined before starting any vocabulary syllabus design. The first priority here, therefore, must be to determine whether we are trying to dictate what learners should be able to or to document what learners can do, i.e. prescribe or describe. The former conforms to the traditional view of syllabus or curriculum development while the latter is closer to the notion of the CEFR. As we have seen, these two views, however, should be complementary to each other, particularly in areas such as multi-word expressions where evidence from learners proves to be difficult to obtain or simply not feasible.

References

Akbrian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. System, 38: 391-401.

Bauer, L. & Nation, I.S.P. (1993). Word families. International Journal of Lexicography, 6(4): 253–279.

Benigno, V. (2012). La notion de collocation fondamentale. Etude de corpus en vue d'une exploitation didactique. Thèse en co-tutelle - Université Stendhal de Grenoble et Université de Palerme.

Boers, F., J., Eyckmans, J., Kappel, J., Stengers, H. & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language Teaching Research, 10: 245-261.

Bonk, W. J. (2001). Testing ESL Learners' Knowledge of Collocations. In Hudson, T. and Brown, J.D. (ed.), A Focus on Language Test Development: Expanding the Language Proficiency Construct Across a Variety of Tests. (Technical Report #21), p. 113-142. Honolulu, University of Hawaii: Second Language Teaching and Curriculum Center.

British Council and EQUAL. (2011). Core Inventory for General English. Accessed http://www.teachingenglish.org.uk/publications/british-council-eaquals-core-inventory-general-english on 19th March 2013.

Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists. English Profile Journal, 1(1): 1-11.

Capel, A. (2013). Completing the English Vocabulary Profile: C1 and C2 vocabulary. English Profile Journal, 3(1): 1-14.

CEFR-J. (2013). New English proficiency guidelines for Japanese EFL learners. Accessed http://www.tufs.ac.jp/ts/personal/tonolab/cefr-j/english/index-e.html on 11th March 2013.

Chen, Y. H. (2011). To wordlist or not to wordlist? The dilemma and challenges for second language learning and testing. In Proceedings of Corpus Linguistics 2011 Conference.

Chen, Y. H. & Baker, P. (2010). Lexical bundles in native and non-native academic writing. Language Learning and Technology. 14(2): 30-49.

Council of Europe. (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Council of Europe.

Council of Europe. (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

Dagut, M. & Laufer, B. (1995). Avoidance of phrasal verbs — A case for contrastive analysis. Studies in Second Language Acquisition, 7(1): 73-79.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? In IRAL, 47, 157-177.

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), Cognition and Second Language Instruction. Cambridge: Cambridge University Press, 33–68.

Gardner, D. (2007). Validating the construct of Word in applied corpus-based vocabulary research: A critical survey. Applied Linguistics, 28(2): 241-265.

Gitsaki, C. (1999). Second language lexical acquisition: A study of the development of collocational knowledge. San Francisco: International Scholars Publications.

Goulden, R., Nation P. & Read, J. (1990). How large can a receptive vocabulary be? Applied Linguistics, 11: 341-363.

Hoey, M. (2005). Lexical Priming. A New Theory of Words and Language. London: Routledge.

Laufer, B. (1998). The Development of passive and active vocabulary in a second language: Same or different? Applied Linguistics, 19(2): 255-271.

Laufer, B. & Paribakht, T. S. (1998). The Relationship between Passive and Active Vocabularies: Effects of Language Learning Context. Language Learning, 48(3): 365-391.

Laufer, B. & Ravenhorst- Kalovski G. C. (2010). Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. Reading in a Foreign Language, 22: 15-30

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners English. Language Learning, 61: 647-672.

Liao, Y. & Fukuya, Y. (2004). Avoidance of phrasal Verbs: The case of Chinese learners of English. Language Learning, 54(2): 193-226.

Milton, J. & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? Canadian Modern Language review, 63(1): 127-147.

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. Bristol: Multilingual Matters.

Nation, I. S. P. (1993). Using dictionaries to estimate vocabulary size; essential, but rarely followed procedures. Language Testing, 10(1): 27-40.

Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In Vocabulary: Description, Acquisition and Pedagogy, N. Schmitt and M. McCarthy (eds), Cambridge University Press, 6-19.

Nation, I. S. P. (2001). Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? The Canadian Modern Language Review, 63(1): 59–81.

Nation, I.S.P. & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31(7): 9-13.

Nation, I. S. P. (2012). Range Program with British National Corpus List 25,000. Accessed http://www.victoria.ac.nz/lals/about/staff/paul-nation on 11th February 2013.

Pearson Education. (2009). The Longman Dictionary of Contemporary English (5th edition). London: Pearson Education.

Read, J. (2000). Assessing Vocabulary. Cambridge: Cambridge University press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Ed.), Vocabulary in a Second Language: Selection, Acquisition and Testing. Amsterdam: Benjamins, 209-227.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. International Journal Studies, 7(2): 105-125.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. Language Learning, 48: 281-317.

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. Language Testing, 18(1): 55–88.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. Language Teaching Research, 12: 329-363.

Schmitt, N. & Schmitt, D. (forthcoming). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Language Teaching.

Van Zeeland, H. & Schmitt, N. (forthcoming). Lexical coverage in L1 and L2 listening comprehension – The same or different from reading comprehension? Applied Linguistics.

Vermeer, A. (2001). Breadth and depth of vocabulary knowledge in relation to L1/L2 acquisition and frequency of input. Applied Linguistics, 22: 217-234.

Zhou, S. (2010). Comparing receptive and productive academic vocabulary knowledge of Chinese EFL learners. Asian Social Science, 6(10): 14-19.

Zechmeister, E.B., Chronis, A.M., Cull, W.L., D'Anna, C.A. & Healy, N.A. (1995). Growth of a functionally important lexicon. Journal of Reading Behavior, 27(2): 201-212.

Giovanna Comerio

The University of Nottingham Ningbo China, Ningbo, China

giovanna.comerio@nottingham.edu.cn

Can the CEFR Assess University Students in a China-Based British University? A case Study at the University of Nottingham Ningbo China

Abstract

The University of Nottingham Ningbo China is an English-medium university that, within the Language Centre, offers its students the opportunity to study an additional foreign language (French, German, Japanese, Spanish) up to the level B1. Despite the introduction of the ELP and gradual alignment of the curriculum to the CEFR, when it was time to align the examinations set up by our University to the CEFR some issues arose. The format of our exam papers is consistent with the programme developed and focus on the intellectual and transferable skills involved. The marking criteria are designed for assessing not only the language, but also the intellectual ability of the students to organize their discourse.

This study is a first step of a reflection on how and to what extent CEFR criteria can be integrated in or replace the HE marking criteria and what kind of changes in the our pedagogy this would involve. As practitioners, this study is a comparison of marks obtained by students taking oral, written, listening and reading test according to the CEFR (DELF A1) and our university marking criteria.

This project would not be possible without the practical help and the reflections of Magali Kerbellec, to whom I am deeply grateful. I wish also to thank Filippo Gilardi for his support.

Short Paper

The motivation

Reflection about teaching has always involved a reflection of the nature and purposes of assessment, consequently over time many analytical or global definitions have been developed by scholars and practitioners. A general definition of assessing could be that it consists in understanding how and what students have learned in terms of knowledge, but also how and to what extent they have developed the skills and abilities connected to that knowledge (Berry 2008). In a constructivist perspective, the language learning being connected with target language culture, assessing language would imply to assess both the usage of the language and the abilities of developing, organizing and negotiate ideas and decisions in the target language. Critical thinking skills are to be expressed in the target language.

The Common European Framework of Languages (CEFR) aims at being an action-oriented "comprehensive, transparent and coherent frame of reference for language learning, teaching and assessment". It defines competences (savoir, savoir-être, savoir faire, savoir apprendre) which are to be realized through tasks where the language is used in a strategic way to achieve a result". The CEFR includes a detailed analysis of descriptors for each level of language competence.

It is important to clarify that in both the constructivist and the CEFR approach, the concept of task is an important one. However, for the former the concept of meaning construction and negotiation underpins the communicative task itself, for the latter the concept of task performed by social actors is related to the solution of a social/cultural problem via the language.

In these two perspectives, the criteria for assessing the language used are different. Holistic and/or analytic are to assess the linguistic performance linked to the intellectual skills: not only the quality of the language is assessed but also how the language is used to achieve the social goals set up by the examinations. For example, they can be expressed in terms of professional skills (i.e. doing a presentation), intellectual skills (i.e. multicultural awareness), transferable skills (i.e. group work).). "Universities need to be assessing the degree to which graduates can display those abilities [general competences]" (Brown, Knight 2004, p. 12).

Critical thinking skills should then be expressed in the target language since, in the words of Whitehead, "the proper function of a university is the imaginative acquisition of knowledge" (1929, p. 145)

One should also be aware that this kind of criteria may pose a challenge in that they might be quite long and discursive, requiring therefore a regular reflection and harmonization between markers.

The CEFR criteria is organized into checklists, detailing the tasks learners should be able to perform in order to effectively co-interact with others.

The question is, do these tasks give universities the information we required? Can they co-operate with others and negotiate meanings by using the target language? Or can they just express information and exchange views? If the professional and transferable skills could be read in terms of savoir-être, than the questions would be how the CEFR links the savoir-être with the savoir and the savoir faire? Does it actually link them?
In the "Referentiel du niveau 1 pour le français", (p. 58) there is a list of expressions the learner should be able to use for "Interagir à propos d'opinions ou de positions" followed by "Interagir à propos d'émotions ou sentiments" (p. 61) and so on. These are however simple verbal or noun phrases, not to be articulated in more complex sentences. The subordinating expressions are detailed in a very short list ("Les connecterus 'logiques' ou argumentatifs", p. 102). Indeed, the CEFR focuses on speech competences, and at A1 and A2 levels they are described mainly as performing personal and basic social relationships, while in higher education institutions, learners are asked to be also able to describe and comment on these, by elaborating the meanings shared and discussed in the taught courses. However, reasoning starts in B1, debating and interacting with an audience in B2.

In creative writing, for A1 and A2 there are no descriptors, B2 learners can express relations between ideas. For the general understanding of oral, only at a B2 level a learner is expected to be able to understand complex information and intervention.
The difference between the nature of learning outcomes in the higher education and the ability to perform social tasks in the CEFR goes in parallel with the difference between the Achievement assessment and the Proficiency assessment, as they are outlined in the CEFR:

Achievement assessment is the assessment of the achievement of specific objectives – assessment of what has been taught. (…) It represents an internal perspective. Proficiency assessment on the other hand is assessment of what someone can do/knows in relation to the application of the subject in the real world. It represents an external perspective. (CEFR, p. 183)

The CEFR also states the difference between using scales and checklists
Rating on a scale: judging that a person is at a particular level or band on a scale made up of a number of such levels or bands.
Rating on a checklist: judging a person in relation to a list of points deemed to be relevant for a particular level or module.
In 'rating on a scale' the emphasis is on placing the person rated on a series of bands. The emphasis is vertical: how far up the scale does he/she come?.(…) The alternative is a checklist, on which the emphasis is on showing that relevant ground has been covered, i.e. the emphasis is horizontal: how much of the content of the module has he/she successfully accomplished? (CEFR, p.189)

Given the differences between the nature of the language tasks and of the assessment, as they seem to be conceived in the Higher Education and in the CEFR, the question is if these are compatible, or can become compatible, and how? Would we lose any important pedagogical features of Higher Education? Would we gain in precision and fairness?
The aim of this study is not to answer these questions now, but rather to compare the two types of assessments and see if any differences exist, and their nature. In particular, in this study, the results of students' exams taken and marked according to the UNNC and DELF format and assessing criteria will be compared.

The context
The University of Nottingham, Ningbo is one of 2 overseas campus of the University (the other being in Malaysia). The three campuses are to deliver highly compatible modules in order to allow students mobility and to ensure a standard of quality. The format of written examination is also the same, the listening and oral examination formats may vary slightly. We also have the same external examiners, whose task is to make sure that the exam formats and marking criteria used are fair and consistent over all three campuses and also conform to British examination standards. The marking criteria for testing productive skills (written and oral) are the same and currently the campuses are working closely together on a new version. We feel that it is not language-centered and needs to focus more on the different levels, but the main issue is on the approach which we would like to take. Do we want checklists, rubrics or scales? If we all agree on accent and pronunciation, vocabulary and grammatical accuracy, our reflection then focuses on content and sentence construction (in particular for beginners). Do we mark on the completion of task or also how it is completed? And what descriptors should we choose for defining this? This debate also includes the proposal for using the CEFR or either just to adopt it, or incorporate it in our marking criteria.

Marking criteria for written work is common to all learning stages, they assess Content, Quality and Range (use of vocabulary and structure) and also Grammatical Accuracy over seven bands ranging from Exceptional (Class I quality) to Hard fail. Exceptional content is: "Extremely well-structured, in-depth coverage of all relevant points plus a high level of original input"; Quality and Range consists of "Extremely sophisticated, complex structures used confidently and fluently. Far beyond (sic) normal expectations at this level"; Grammatical Accuracy consists of "The complexity of the language is matched by extremely accurate usage and excellent grammatical awareness. No errors".

Besides the difficulties that one might have when using these criteria for beginners, the interesting point here is the content and the quality and range: we teach our students that they need to express their own personality in the target language.

The assessment of oral skills are supported by 2 sets of marking criteria, one for stage 1 (beginners) and the other for stage 2 and higher.

Stage 1 marking criteria assess four areas: Communication and Understanding including Completion of Tasks, Grammar, Range of Expression and Linguistic Structures, Accent and Pronunciation. Class I performances would be described as follows regarding

Communication and Understanding: "Full and active participation; very effective communication; excellent level of understanding, few lapses; task completed". For stage 2 and above, assessment includes Accent and Pronunciation; Grammar, Vocabulary Register, Linguistic structures, and Range of expression; Intellectual Performance, Knowledge, Conceptual Grasp, Ability to Sustain Argument, Analysis, Originality, Discursive Organization and/or, where relevant, Completion of Task. Overall, we praise the knowledge gained through independent study, and we assess the fluency, say the ability to effectively convey meanings without loss of clarity due to accent and pronunciation, grammatical or syntax mistakes. We seem to mainly assess and aim to a fluent communication, where for fluency we intend "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" (Lennon, in Riggenbach 2000, p. 25).
Marking essays and oral performances is a challenging task and tutors need to harmonize their marking. Orals are marked by two examiners working together and discussing their blind marking. Essay marking is harmonized by initially sharing and discussing the marking of one third of the papers.

It is not surprising that the main areas for discussion are the evaluation of the content, quality and the range. For example, as the CEFR in A1 level lists linking devices such as "et, mais, alors, parce que, pour, après" we expect students to use them at any time when the occasion arises. Vocabulary and linking devices should fit in a coherent discourse appropriate to the situation. The level of the "quality and range" is therefore deeply connected with originality and structure of the content.

Putting aside the consideration about long titles descriptors, their organisation, and even the details of these descriptors, what is more important to notice here is the emphasis on the quality of participation, the communication on one side, and the quality of the content and its organisation on the other. These four elements are consistent with the very purposes of higher education mentioned above.

The subjects
The majority of the university students are Mandarin native speakers, studying in English. After a preliminary year where they focused on the acquisition of English for Academic Purposes, they start a 3-year academic course, where many of them choose to add a language as a degree component, while others choose to study the language as an optional modules. Compulsory modules are worth 40 credits per year, which are organized over four taught hours per week plus 6 hours of self-study. Our university, semester A lasting 12 weeks, and semester B 11 weeks, learners have therefore 48/44 contact hours per semester (92 per year), and 72/66 hours of self-study (138 per year) for a yearly total of 230 hours. As a university, we strongly believe in autonomous learning strategy development, therefore we added for the first 2 years of study one extra contact-hour called "self-study hour" where students learn how to improve their language learning strategies.

Many of our students will not have direct contact with the target language countries except via the Internet, therefore all language teams developed extra-curricular activities for enhancing students motivation and engagement. Around 10 to 15% of the students attend regularly extra-curricular activities.

The population of this study includes 11 undergraduate students of year 2 (year 1 academic), 2 males and 9 females. All are French beginners having studied French for one semester (48 hours, 72 self-study), who volunteered to participate in this study. None of them had any experience of DELF examinations or had any training.

A comparison of the examination results
Since the marks of the written, oral, listening and reading skills were differently weighted the following standardization was made:

- the DELF written task worth 25 points, 50 for UNNC, it has been standardized at 50 points;
- the DELF listening task worth 25 points, 100 for UNNC, it has been standardized at 100;
- the DELF oral task worth 25 points, 100 for UNNC, it has been standardized at 100;
- the reading task in both cases are worth 25 marks.

a) The reading comprehension task

The DELF reading comprehension passages are short texts such as an invitation to a party and information from which students should extract simple information. The test consists of 4 tasks to be completed in 30 minutes. All the marks range between 15 and 25, with 7 students being scored between 21 and 24. The tasks are multiple choice questions, fill in the blanks and open-ended questions.

The UNNC test has to be a passage of about an A4 page, usually a biography, an interview, a description of a place, or a text related to a particular aspect of culture and traditions. Students are usually asked to answer open-ended questions (but not in this paper, because we wanted to make it as similar as possible to the DELF paper), multiple choice questions and True or False questions with the need for students to justify their answer by quoting the sentence from where they extracted the answer. If students don't justify their answers they get no mark for the full sentence; if the answer is right but the justification is wrong or vice versa, the mark will also be zero. The aim of this format is that we want students to analyze and process the information correctly. We usually suggest that students devote approximately 30 minutes to the completion of this task. The marking range was from 14.5 to 24, with two students awarded 6.5 and 7.5 marks because they didn't justify any of their answers in the True or False questions.

If we compare the marks, the DELF average is 20.8 and UNNC average is 18.8. The DELF mark fell in the upper Class I quality (83.2/100), the UNNC average falls in the lower Class I quality (75.2/100). Only one student falls from one band with the UNNC test, all the others remain in the same band, the only variation being the upper level of it.

Out of 25 points

Student	DELF	UNNC	Class
1	24	24	EI, EI
2	23	21	EI, I
3	23	6 (16)	EI, II.1
4	20	21.5	Iu, Iu
5	21	17.5	Iu, Il
6	22	17.5	Iu, Il
7	19	22	Il, Iu
8	20	15.5	Iu, II.1
9	21	19	Iu, Il
10	21	7.5 (17.5)	Iu, Il
11	15	14.5	II.1, II,1

b) The listening task

The DELF test lasts 20 minutes and includes 3 tasks, each one repeated twice for a total of 12 questions aimed at testing the detailed and global comprehension through ten multiple choice questions and 2 fill-in-the-blank questions. Also in this test, students scored impressive results, ranging from 48/100 to 96/100.

The UNNC tests consists of 30 multiple choice questions, organized into 6 or 7 tasks, each one repeated three times, total duration being 45 minutes. The results were impressive, with marks ranging from 63.3 to 90/50.

The average of DELF scores is 74.9/100 (I class), the average of UNNC scores is 72.3 (I class). The average for the two exams did not differ much, however individual student performances are very different and somewhat surprising. This is the first test where some students out performed in the UNNC test than the DELF test: 5 out of 11 students improved their positions by one (3 cases) or two (2 cases) classes. Four EI scores fell by one class, and two students stay in the same class.

Out of 100

Student	**DELF**	**UNNC**	**Class**
1	100	86.6	EI, I
2	92	90	EI, EI
3	80	76.6	I, I
4	96	80	EI, I
5	96	76.6	EI, I
6	48	86.6	III, I
7	48	76.7	III, I
8	56	76.6	II.2, I
9	64	76.6	II.1, I
10	96	83.3	EI, I
11	48	63.3	III, II.1

c) The written task

The written task of the DELF lasts 30 minutes and includes 2 tests: filling in a form and writing an email on an everyday life topic. Our students' marks range from 32/50 to 42/50, except for one case who received 28 marks.

The UNNC written task has to be completed in two hours and includes: reading comprehension, a grammar section, and writing tasks. Students are usually trained in completing the written task in approximately one hour and many of them choose to write it first. Students have to write a 150 word essay from a choose of two topics: in this case they had to choose between describing a person they know or writing a postcard (letter) about their holidays. UNNC marks range from 25.1/50 to 37/50.

Out of 50 points

Student	**DELF**	**UNNC**	**Class**
1	42	34	I, II.1
2	42	37	I, I
3	36	28	I, II.2
4	41	34	I, II.1
5	47	32.1	I, II.1
6	28	25.1	II.2, II.2
7	39	26.3	I, II.2
8	36	32	I, II.1
9	35	29.3	I, II.1 (borderline)
10	34	30.8	II.1, II, 1
11	32	28.8	II.1, II.2

The difference between the students' performance is significant: the DELF average is 37.4/50 (74/100), the UNNC is 30.6/50 (61.3/100). According to the UK score ranking, DELF examination results are to be put in class I quality, while UNNC results are class II.1. With the DELF, 8 out of 11 papers are Class I, two in class II.1 and one in class II.2. According to the UNNC only one paper could be ranked in class I, 5 in class II.1, and four in class II.2. Three students scores are in the same band (n. 2, n. 6., n. 10) for both the DELF and UNNC, the other eight students fall in an higher band with the DELF, and fall by one (6 cases) or two (2 cases) bands with the UNNC scores.

d) The oral task

The DELF oral examination consists of three individual exchanges with the examiner (the examiner asks questions, the examinee ask questions to the examiner using the clue-word given, a role-play), in total it lasts for 5–7 minutes with 10 minutes of preparation time for the role-play. Students' performance is impressively good, and marks range from 82 to 100. UNNC oral test consists in a very short real of fictional coherent self-introduction according to three requirements and lasting 30–60 seconds, and a conversation of 4–5 minutes between three students on a given topic whose requirements are specified. The conversation needs to include this required information, but should also go behind it. The marks awarded for the oral test range from 56.7 to 68.5.

Out of 100 points

Student	DELF	UNNC	Class
1	92	63.5	EI, II.1
2	100	68.2	EI, II.1
3	94	56.7	EI, II.2
4	92	63.7	EI, II.2
5	100	60	EI, II.2
6	69	61	II.1, II,1
7	90	68.5	EI, II.1
8	80	61.5	I, II.1
9	88	67.5	I, II.1
10	82	67	I, II.1
11		61	

The oral test DELF average is 88.7/100 (upper class I), the UNNC average is II.1 class (63.7/100). It is interesting to notice that 3 of the 6 DELF exceptional I class fell two classes and 3 fell three classes.
The three DELF I classes fell one class, and only one student gained the same class score in both of the exams.

Provisional conclusions and next steps

This study is just the first, partial phase of an analysis needed for understanding if and to what extent the CEFR can be used in higher education institutions.
The results of the written tasks seem to show that there is a significant difference between the students' ability of writing short or longer texts about themselves. On studying the examiners' analysis it appears to emphasize the relevance of their appreciation of syntax and the quality of the content. While there are no discrepancies in assessing the grammatical accuracy, the emphasis on text cohesion, coherence and meaning makes the difference. The UNNC marking criteria stresses the usage of linking devices on one hand, and the originality and logical structure of the texts produced on the other. We will need to compare in greater detail the marks awarded according to each component of both the DELF and UNNC descriptors and also to use the DELF criteria for marking the UNNC papers and vice versa.

The similar results of the two types of reading comprehension show that, in contradiction with our expectations, the nature and length of the tasks and the type of the questions did not impact as much on the reading performance. It is true that True or False questions with justification, was the most challenging for students; nevertheless, this seems to have been balanced by the DELF open-ended questions.

As far as the oral performance is concerned the difference between the DELF scores and the UNNC scores are striking: performance according to the DELF format is upper class I, while the performance according to the UNNC format is II.1 class. Nine out of 10 from the DELF to UNNC exams fell by one or two classes. According to the DELF criteria the range of vocabulary and the quality of the grammatical accuracy are excellent or very

good, while the orals marked according to the UNNC criteria seem to include a series of mistakes in interrogative sentence construction, usage of preposition, and an influence from the English language (e.g. dates and age). Likewise for the written, the UNNC oral marks are shaped by the emphasis put on the quality of interaction and communication, the completion of the task Also in this case it would be interesting to compare the scores gained for each component of the marking criteria and to mark the UNNC performance with the DELF criteria and vice versa.

The results of the receptive skills are both surprising and contradictory. Although the reading tasks are very different in nature, length and in part the type of questions and their rationale, both the global and the individual students' results are very similar. An investigation for the reasons behind this similarity is however needed, in order to make sure that it is not just a lucky coincidence.

Finally, the results of the listening examination are equally surprising and also contradictory. They are contradictory because they show a similar average (as do the reading exams), but the individual marks are very different. They are also surprising because some students scored very low in the DELF but very high in the UNNC test, which was not the case for the other tests.

The first results of the productive skills performance seem to confirm that the difference between the approaches and aims of the DELF and our university as they are reflected in the marking criteria heavily impacted on the performance assessment and appreciation. Making the CEFR and UNNC compatible would therefore involve a deep reflection and a radical change of pedagogical perspective and teaching/learning aims. We do not have to forget that the CEFR criteria are to be used to certify the language level reached by a learner in different learning environments and from different experiences. In universities we aim at support learners to become independent, reflective and knowledgeable (Brown, Knight, 2004, p.54). The language tasks are therefore to be accomplished in a social and engaging environment where some level of divergent thinking is stimulated. "There is some evidence that divergent tasks are more prone to communication breakdowns within the groups, but they also yield greater learning than simpler tasks" (Weissberg, 2001).

References

Beacco, J. C. (2007). L'approche par compétences dans l'enseignement des langues. Enseigner a partir du 'Cadre européen commun de référence pour les langues. Paris : Didier.

Berry, R. (2008). Assessment for Learning. Hong Kong: Hong Kong University Press.

Biggs J. & Tang C., (2011). Teaching for Quality Learning at University (4th edition. Maidenhead: McGrawHill

Boud, D. (1990). Assessment and the Promotion of Academic Values. Studies in Higher Education. 15:101-111

Brown, S. & Knight P. (2004). Assessing Learners in Higher Education. London-New York: RoutledgeFalmer.

Campbell, A. & Norton, L. (Eds) (2007). Learning, Teaching and Assessing in Higher Education. Developing Reflective Practice. Exeter: LearningMatters.

Conseil de l'Europe, Division des Politiques linguistiques. (2007). Niveau A1 pour le français. Un référentiel, Paris : Didier.

Council of Europe. (2001). The Common European Framework of Reference for Languages. Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Genesee, F.& Upshur, J.A. (1996). Classroom-based Evaluation in Second Language Education. Cambridge: Cambridge University Press.

Nunan, D. (1989). Designing Tasks for the Communicative Classroom. Cambridge: Cambridge University Press.

Puren, C. (2010). Entre l'approche communicative et la perspective actionnelle, quoi de neuf? Les cahiers pédagogiques. Hors-série numériques, 18: 87-91.

Ramsden, P. (2003). Learning to teach in Higher Education (2nd edition). London: Routledge.

Richards, J.K. & Lockhart, C. (1996). Reflective Teaching in Second Language Classrooms. Cambridge: Cambridge University Press.

Richer, J.J. (2009). La compétence ou deux lectures divergentes possibles du 'Cadre européen commun de référence pour les langues'. Les cahiers de l'Asdifle, 20 : 184-208.

Riggenbach, H. (2000). Perspectives on Fluency. Ann Arbor: The University of Michigan.

Rost, M. (2002). Teaching and Researching Listening. London: Pearson.

Ur, P. (1984). Teaching Listening Comprehension. Cambridge: Cambridge University Press.

Weissberg, R. (2001). Talking to Learn. Socializing the Language Classroom. Bräuer, G. Pedagogy of Language Learning in Higher Education.

Whitehead, A. N. (1929). The Aims of Education and Other Essays. Cited in Ramsden, P. (2003) Learning to teach in Higher Education (2nd edition, p. 21). London: Routledge.

Michael Corrigan

University of Bedfordshire, Cambridge, United Kingdom

corrigan.m@cambridgeenglish.org

Interchangeability of Test Results and the CEFR – a Validity Argument Approach

Bio data

Michael Corrigan is in the process of completing a PhD in language testing at the University of Bedfordshire. His presentation relates to part of this project. Until recently, he worked with several members of the Association of Language Testers in Europe (ALTE) on the (mainly statistical) validation of their tests. This work provided both data and inspiration for his PhD. His work now involves Cambridge English Language Assessment tests.

Abstract

Descriptors of the CEFR are commonly provided to test users to assist them in interpreting the test results. However, when comparing the results of different tests targeted at the same CEFR level, users may be tempted to treat them as interchangeable because the descriptors given are identical. Issues of test-CEFR alignment notwithstanding, the CEFR alone does not support such interpretations, as, according to the Manual on Relating Examinations to the CEFR (Council of Europe, 2009:4), 'two examinations may both be "at B2 level" and yet differ considerably'. How, then, can we determine whether the results of two exams at the same CEFR level may be used interchangeably? Consideration of this question requires an appropriate methodological approach which takes the intended use of results into account, as they may be interchangeable for some purposes and not others. Although a number of methodologies to appraise the link between tests have been put forward (e.g. Mislevy, 1992; Linn, 1993; Kolen and Brennan, 2004), none are entirely satisfactory. An approach based on validity arguments (e.g. Kane, 2012) will be outlined in this presentation, along with the results of a small study investigating the comparability of two exams of different foreign languages: Cambridge English First (FCE) and Certificato di conoscenza della lingua italiana 3 (CELI 3), an exam of Italian. This talk will be of interest to those seeking to link their tests to the CEFR and test users who must interpret test results for a range of different tests.

Short paper

Making decisions based on test results is a highly problematic area, not least because those making the decisions do not always know enough about testing in general, or the particular tests they are concerned with. This applies to the use of a single test but the problems are compounded when, as is sometimes the case, results from any one of a number of tests is accepted. Examples include tests used to make decisions concerning employment, for immigration purposes and those used for some educational purposes. The CEFR is becoming increasingly employed in such situations as a guarantee of interchangeability. Bonnet (2007) provides such an example. The French Department of Education, in order 'to revitalize the teaching and learning of FLs', introduced foreign language competency targets for learners in which targets were specified in terms of CEFR levels. These were A1 at the end of primary school, B1 at the end of compulsory

education, B2 at the end of upper-secondary school, or C1 for those attending special language classes. These targets were introduced in 2005 and coincided with two existing performance targets for the Department, based on the proportion of those achieving foreign language competence at A1 at the end of primary and B1 at age 15. Special exams were commissioned by the French government from various test providers, and offered to learners on a voluntary basis, first in German, then in English and Spanish, with the intention to follow with other languages at a later date. The French Government clearly thought that the results of the tests they commissioned were sufficiently interchangeable for the purposes of their policy, however, it is not clear from Bonnet's paper whether this decision was based solely on the posited CEFR level of the tests, or involved other research.

Using the CEFR as an instrument to underwrite interchangeability of test results may problematic depending on the purpose of the comparison. Such interchangeability would, presumably, rest on the test results correctly indicating which CEFR Can Do Statements may be applied to specific candidates. However, according to the Manual for Relating Examinations to the CEFR, 'There is no suggestion that different examinations that have been linked to the CEFR…could be considered to be in some way equivalent. Examinations vary in content and style…so two examinations may both be "at B2 level" and yet differ considerably' (Council of Europe, 2009:4). In an empirical study examining the relationship between French, German and English tests and the CEFR, Noijons and Kuijper (2006, 2010) find much the same: broad similarities, such as increases in cognitive and linguistic complexity which related to those suggested by the CEFR and test-specific differences, including lack of correspondence to some CEFR descriptors and variation in text and task types also featured. Such divergence between tests should not be a surprise where tests are not developed in parallel, with the aim of yielding interchangeable results, such as those of SurveyLang (2012). Neither, of course, was the CEFR designed as a tool to capture fine-grained similarities and differences between tests. As Milanovic (2009) points out, it is intentionally underspecified in order to avoid it being context-specific, and thereby less broadly applicable. As Coste (2007) explains, the CEFR, particularly in the case of assessment, is used in ways for which it was not originally designed.

Where interchangeability of test results is important and the CEFR is inadequate for the intended use, further research must be undertaken to substantiate direct link between the two tests. Research methods for such linking will not be discussed here, but rather the framework in which the results of such research may be understood. Similar frameworks developed by Mislevy (1992) and Linn (1993) are based on a taxonomy of types of linking, where a number of approaches to linking, including data requirements, methods and permissible uses of results, are described. Together, these types of linking form a continuum, which goes from those designated as strong forms of linking, such as equating, to weaker forms, such as social moderation. To understand the nature of a linking project using the Mislevy/Linn approach, it would only seem necessary to locate your linking on the continuum in order to understand, relative to the other linkings, the implications of the project. However, as Newton (2010:41) points out, such taxonomies flatter to deceive: they do not really form continua which can be used in this way: 'Exactly what distinguishes linking relationships at different points along the continuum is not always clear', it may be 'the idea of strength…expressed in terms of methodological rigour', or it may be the 'extent to which key assumptions have been satisfied'.
Kolen and Brennan (2004) offer an alternative approach: degrees of similarity. In their view, 'the utility and reasonableness of any linking depends upon the degree to which test share common features' (p434). To capture the degree of similarity between tests, they nominate four general areas for investigation: inferences (from test results about candidates), constructs, populations and measurement characteristics/conditions (including a diverse range of characteristics, such as test specifications, reliability and conditions of administration, which, in Generalizability Theory are considered facets of measurement – features which may vary between administrations or test forms and

therefore lead to variation in test results). Little guidance, however, is offered on how this scheme might be put into practice. Furthermore, in the case of tests of different foreign languages, some thought would need to be given as to how constructs and populations in particular would be compared.

The approach suggested in the current paper is based on the validity argument approach to validation (Kane, 2006, 2012). In Kane's scheme, the recommended interpretations of test results, and therefore uses, must be detailed and supported by an argument containing theory and evidence. Such an approach has the benefit of focussing all the work concerning validity onto the appropriateness of the use of the test results. The rationale is to ensure that, if some aspect of the testing process does not permit the use of the test results for a particular purpose, it will be prevented and not obscured by mountains of uncoordinated research. In the case of different tests being used interchangeably, the relevant recommended interpretations and uses of the results of each test must match those of the others. As a consequence, those interested in using test results interchangeably must focus their attention on the recommended interpretations of test results and the validity arguments which support them and then compare them across tests.

In order to make a decision concerning a candidate based on their test result, the CEFR may be used as an aid to interpretation. The test result may be described in terms of Can Do Statements, implying that the candidate has a good chance of being able to do what is suggested at least adequately and at least most of the time it is called for. The use of a Can Do Statement cannot imply more than this, however. Among other things, the Statements lack specificity on matters like range of what a learner can do and quality with which he or she can do them (Green, 2010; Hulstijn, 2007), contain a number of anomalies and inconsistencies (Alderson et al., 2006), display a limited treatment of contextual variables (Weir, 2005) and do not form a scale which is based on a theoretically-grounded construct of language proficiency (Fulcher, 2004).

Considering the limitations of the Can Do Statements, why should anyone wish to use them as a basis for treating test results as interchangeable? One reason might be that they are, nevertheless, sufficiently representative of what is required: a general level of proficiency which need not be more specific. For example, in Bonnet's example of the French education system, perhaps that is enough, given that the results are used as performance targets for an educational system and department. In other cases, such as in recruitment, there is an opportunity to think more carefully about the language use which may be required of the job holder. Even in the case of migration and citizenship-related tests, what is tested can be related to contexts of language use (Balch et al., 2008; Gysen & Van Avermaet, 2005). In fact, the more specific the intended context of language use, the more likely the CEFR alone is inadequate to guarantee sufficient interchangeability.

Test users will not always have the time or inclination to find out more about language tests in order to determine whether they are suitable for the intended purpose. In these cases, suitable test use always boils down to adequate assessment literacy. Test providers are considered to have the responsibility to do as much as they can to inform users of appropriate interpretations for test results (AERA, APA, & NCME, 1999; Association of Language Testers in Europe, 2010; European Association for Language Testing and Assessment, 2006). Such efforts are easier when test use applies to more specific context, however, as there is more scope to assist users in understanding the recommended interpretations for their contexts. IELTS, for example, produce materials explaining test results for groups such as admissions tutors (IELTS, N.D.). At the same time, it is also true that, the more specific the context, the less useful the CEFR is in providing a link between tests because its descriptors are intended to be context-neutral.

References

AERA, APA, & NCME. (1999). Standards for Educational and Psychological Testing (2nd ed.). Washington, DC: AERA.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. Language Assessment Quarterly, 3(1), 3-30.

Association of Language Testers in Europe. (2010). The ALTE Code of Practice. Retrieved from http://www.alte.org/downloads/index.php?docid=167.

Balch, A., Corrigan, M., Gysen, S., Kuijper, H., Perlmann-Balme, M., Roppe, S. & Zeidler, B. (2008). Language tests for social cohesion and citizenship – an outline for policymakers. Paper presented at the Linguistic integration of Adult Migrants, Strasbourg. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/ALTE_migrants08_final_EN.doc.

Bonnet, G. (2007). The CEFR and Education Policies in Europe. The Modern Language Journal, 91(4), 669-672.

Coste, D. (2007). Contextualising uses of the Common European Framework of Reference for Languages. In F. Goullier (Ed.), The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities - Intergovernmental Language Policy Forum. Strasbourg: Council of Europe.

Council of Europe. (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) - A Manual. Retrieved from http://www.coe.int/t/dg4/linguistic/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf

European Association for Language Testing and Assessment. (2006). EALTA Guidelines for Good Practice in Language Testing and Assessment Retrieved from http://www.ealta.eu.org/guidelines.htm

Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. Language Assessment Quarterly, 1(4), 253-266.
Green, A. B. (2010). Requirements for Reference Level Descriptions for English. English Profile Journal, 1(01), 1-19.

Gysen, S. & Van Avermaet, P. (2005). Issues in Functional Language Performance Assessment: The Case of the Certificate Dutch as a Foreign Language. Language Assessment Quarterly, 2(1), 51-68.

Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency1. The Modern Language Journal, 91(4), 663-667.

IELTS. (N.D.). IELTS Scores Explained: IELTS.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (4th ed.). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2012). Validating score interpretations and uses. Language Testing, 29(1), 3-17.

Kolen, M. J. & Brennan, R. L. (2004). Test Equating, Scaling, and Linking: Methods and Practices (2 ed.). NY, NY: Springer.

Linn, R. L. (1993). Linking Results of Distinct Assessments. Applied Measurement in Education, 6(1), 83 - 102.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. Research Notes, 37, 2-5.

Mislevy, R. J. (1992). Linking Educational Assessments: Concepts, Issues, Methods, and Prospects Policy and Research Reports: ETS.

Newton, P. E. (2010). Thinking About Linking. Measurement: Interdisciplinary Research & Perspective, 8(1), 38 - 56.

Noijons, J. & Kuijper, H. (2006). Mapping the Dutch Foreign Language State Examinations onto the Common European Framework of Reference - Report of a Cito research project commissioned by the Dutch Ministry of Education, Culture and Science. Arnhem: CITO.

Noijons, J. & Kuijper, H. (2010). Mapping the Dutch Foreign Language State Examinations onto the CEFR. In W. Martyniuk (Ed.), Aligning Tests with the CEFR - Reflections on using the Council of Europe's draft Manual. Cambridge: CUP.

SurveyLang. (2012). First European Survey on Language Competences: Final Report Version 4.0. Cambridge.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing, 22(3), 281-300.

Lieve De Wachter & Jordi Heeren

Katholieke Universiteit Leuven, Leuven, Belgium

lieve.dewachter@ilt.kuleuven.be

Can a Language Test Verify the Academic Literacy of University Students and how Does that Relate to Study Success?

Bio data

Prof. dr. Lieve De Wachter teaches Dutch applied linguistics and Dutch for academic purposes at the university of Leuven. She teaches academic writing and presentational skills at the faculties of Social Sciences and Arts. She is promotor of the project TaalVaST (entry-level language skills). In the project a language test, a digital learning environment and a series of language workshops are being developed.

Jordi Heeren works as a project assistant for TaalVaST. He is also responsible for the development of language supporting initiatives such as the digital learning environment and the language workshops.

Abstract

The TaalVaST language test, developed at the Leuven Language Institute (University of Leuven) has been taken by more than 9000 first year students since September 2010. Its growing popularity shows that a well constructed and piloted language test can replace other screening methods that take up a lot of time and resources. The test has proved to be an efficient tool to, on the one hand, give students an early 'warning signal' during their academic education and on the other hand function as a starting point for language tutoring tailored to the needs of a defined target audience.

The construct of the language test has been proved to be statistically valid and reliable. A pilot version of the test was used to determine the construct validity and reliability by means of the simple-item discrimination method, the calculation of the point-biserial correlation coefficient, item facility and the Kuder-Richardson 20 formula. By analyzing students' secondary education and language use at home the group-differential validity of the test has been determined. Moreover, a study by Huyghe and Marx (2011) found a significant correlation between the language test scores and students' academic success. That confirms other important studies in the field of language testing (Van Dyk 2010, McNamara 1996) and proves the concurrent validity of the test (Davies 1990, 23-24).

Constructing the test, we deliberately did not start from the CEFR, but used a needs analysis as a basis for the construction of our test items. Essentially, the reading and writing requirements for first year university students were taken as a basis for the items and were only incorporated into the language test after they had proven to be valid. It is demonstrated that the development of a language test is a continuous process of designing, testing and revising.

Short paper

Introduction

Academic bachelor programs at Belgian universities and in this article more particularly at the University of Leuven are confronted with a low academic success rate of first year students: only 41% of the students passed all their exams in 2012 [1]. One of the reasons often mentioned for this phenomenon is that higher education in Belgium is open to all students, regardless of their study or grades in secondary school. Moreover, there is no standardized test at the end of secondary education or prior to university studies as is for example the case in the Netherlands and in most of the Anglo-Saxon world. That is why the intake of first-year university students has become much more varied during the last two decades (De Wachter 2010; Peters et al. 2010).

De Wachter & Cuppens (2010) make it clear that, because of that growing diversity, not all students are equally well prepared for a university education. Whereas academic bachelor programs typically require a strong academic preparation in study skills and content, many students lack both these skills and the necessary content knowledge (depending on the faculty the content will be specified as mathematics, life science, social science, languages etcetera). The coaching programs and summer schools organized by several faculties focus mostly on content development and processing.

In academic research on the subject, several retention studies focus on cognitive variables such as high school GPA as typical predictors of study success (Tyson 2011, Veenstra et al. 2008 & Zhang et al. 2004), while motivational characteristics, such as the level of autonomous motivation and academic self-concept, have been repeatedly associated with academic achievement as well (Guay et al. 2008, Marsh & Craven 2006). This article will bring up yet another powerful predictor of study success: academic language skills or, even broader, academic literacy (Brown & Hudson 2002, De Wachter & Heeren 2013, Van Dyk 2010, Holder et al. 1999, Marx & Huyghe 2011, Peters, Van Houtven & Morabit 2010). As Peters and Van Houtven (2010, 16) suggest, language problems (can) eventually result in study problems.

In the following paragraphs the conceptual set-up of an academic literacy test and its results will be discussed, referring mainly to a correlation study that links the academic language test results to early academic achievement. Moreover, the implications of these results for academic coaching and practice of first year students will be briefly mentioned.

Development of a test of academic literacy

Test purpose and content

One of the crucial steps in test development is defining its purpose. The main goal of the academic language test discussed in this paper is to give students an early warning signal so that steps can be taken at an early stage to remedy their deficiencies or to reorient them if necessary. That implies that the test is informative and hence rather low-stakes, but that it also wants to show students their actual academic language proficiency and to a certain extent their 'academic potential'. In order to convey the correct message to the students, the validity and reliability of the test have to be thoroughly investigated. Moreover, because it is not only important to identify but also subsequently assist those at-risk students, it must be examined whether the outcome of the test is meaningful so that it serves its assumed purpose.

Inherently connected to the test purpose is a clear definition of the target language use and the accompanying language tasks. Important is that the test content has initially not been determined by the Common European Framework (CEFR). Instead, a needs analysis

[1] www.kuleuven.be/toekomstigestudenten/studievoortgang.html

has been the starting point to find the required language proficiency level and specific language tasks. The outcome was comparable to Van den Branden's definition of academic language use (Van den Branden 2010, 216-217). He characterizes academic language as having a high amount of non-frequent vocabulary, complex grammatical structures and impersonal language with implicit relations between text parts. Hence, the language test has been developed to test students' (meta)linguistic strategies. These consist of, for example, reading, inferring meaning from context and proving insight in text structure, rather than more 'elementary' language aspects such as spelling and (basic) grammar. The findings of our needs analysis were compared to the "Startcompetenties Hoger Onderwijs" developed by the SLO, which are competency-levels for mother-tongue speakers of Dutch based on the CEFR (Bonset & de Vries 2009). Our analysis shows that the language tasks required of the students can be situated often at a C1-level.

Since the test tasks should preferably reflect actual tasks in real language situations, only authentic materials such as fragments from first year syllabi or handbooks have been used (Bachman & Palmer 1996, 11). That authenticity will affect the test-takers' perception of the test and stimulate a positive affective response to the test (Bachman & Palmer 1996, 24). To ensure that the texts used in the test reflect the intended level of complexity, the Flesch-Douma readability index was used (Jansen & Lentz 2008, Hacquebord & Lenting-Haan 2012). In addition, a word frequency tool determined the word frequency of the word-knowledge items in the test (Hazenberg & Hulstijn 1996). Based on text complexity and word frequency the test can also be classified as a C1-language test. A more important aspect of the test however is that it does not only want to measure a certain level of language mastery but also a much more general concept of academic literacy or even academic potential. In other words: it wants to measure whether students have the language strategies needed on an academic level.

Test type

An important aspect of the test design is the type of test that is to be created. Two main categories can be discerned although other options that contain elements of both are possible. A test can be classified on the one hand as an achievement test (testing a certain amount of knowledge or skills acquired in a specific period of time) or, on the other hand, as a proficiency test (McNamara 2000, 6-7; Davies et al. 1999). The academic language test developed at the Leuven language institute is an example of a proficiency test because it "look[s] to the future situation of language use without necessarily any reference to the previous process of teaching" (McNamara 2000, 7). Brown and Hudson (2002) also mention the difference between criterion-referenced tests, that look towards students mastery of several clearly defined tasks or goals, and norm-referenced tests that are "primarily designed to disperse the performances of students in a normal distribution based on their general abilities, or proficiencies" (Brown & Hudson 2002, 2). The academic literacy test is clearly the latter.

Practicality of the test

The form and practicality are other aspects to consider when creating a useful test (Bachman & Palmer 1996, 18). The Leuven language test is a digital test, which ensures the feasibility of the organisation of the test, its correction and the processing of the data (McNamara 2000, 80). For example, it enables us to install a time limit of 30 minutes after which the test automatically ends and the results are immediately processed and shown to the students. There are however some constraints inherent to the computer medium as well. An extensive production response such as free writing, for example, cannot be used, only selected and limited production responses are possible (Bachman & Palmer 1996, 55). Sercu et al. (2003, 109) however believe that even with these restraints, a well-considered set of questions can still render reliable results.

The initial test design proves to be a crucial step in the testing process. It helps to get a clear view on several fundamental aspects such as test purpose and test content.

The pilot version

Every proficiency test, however low-stakes, asks for a pilot. The pilot is used to check the validity and the reliability of the test using an audience that approximates the intended target audience. In this case, the pilot consisted of 652 participants, 302 of whom were last-year secondary students, 336 first year university students and a small group of 14 second language learners (L2). The L2-learners already had a B2 degree and were at the time following a course of academic Dutch in order to obtain a C1 certificate. 191 of the 302 secondary school students followed general secondary education (ASO in Flemish school system) preparing students for a study in higher education; the other 111 secondary school students followed a technical education (TSO). More and more often, these students also start in higher education, often in a professional bachelor program, although that is not always the primary focus of their previous education.

To ensure the defensibility and fairness of interpretations based on the test results, different fundamental measurements of validity and reliability have been used. Firstly, the simple item discrimination and point-biserial correlation coefficient have determined the content validity of the test items (Brown & Hudson 2002, 118, 130). The validity of test items indicates whether they measure what they intend to measure. Items that proved invalid have been left out; eventually 25 items have been selected. Secondly, the reliability of the test has been calculated. Reliability is concerned with the consistency of measurement, regardless of the characteristics of the actual testing situation. In our case, it has been measured using the Kuder-Richardson formula 20 (Davies 1999, 22). To have a reliable test, the outcome of the formula should be higher than 0.70. The Leuven test scored 0.78 and proved to be reliable, considering its low-stakes purpose.

Besides the reliability and the content validity, the differential-groups construct validity of the academic literacy test has also been investigated (Brown & Hudson 2002, 230-233). The following graph shows the scores of the different groups that participated in the pilot.

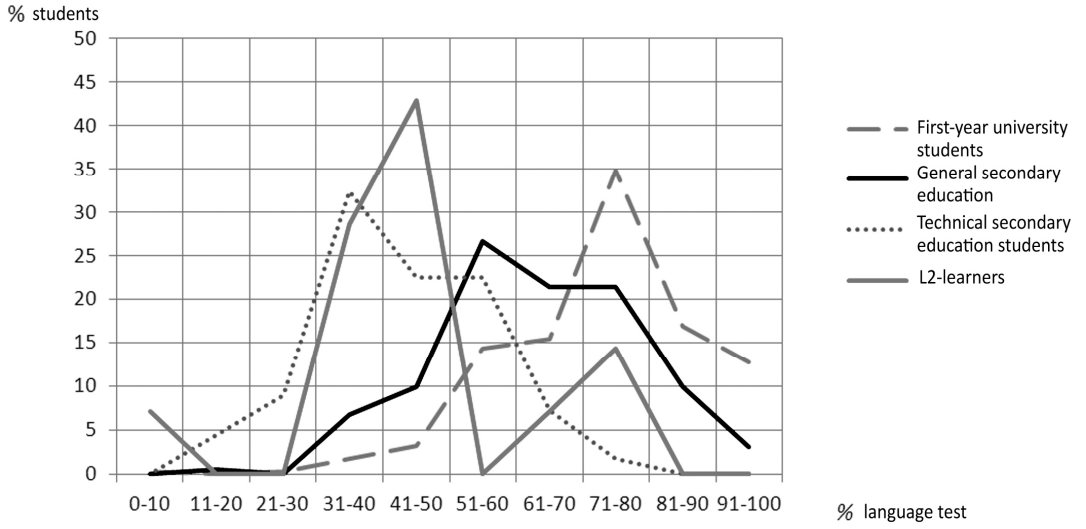


Fig. 1: pilot results

Although there are overlapping areas, the average test scores of the subgroups differ significantly. The highest scoring group is the group of first-year bachelors. These students had been studying at university for several months already and they scored differently from the students in general secondary education (n=191). In general, the latter group is prepared for a study in higher education, but not all students necessarily start a university education. In technical secondary education (n=111), students are mainly prepared for technical functions or for a professional bachelor. Depending on the school and on the program they followed these students sometimes register for a

university education as well. The position of the L2-learners is more difficult since their group is too small to claim general tendencies, although their average score is quite low.

External validity: test results and correlation with exam scores

In September 2010, after the 2009 pilot, the academic literacy test was for the first time taken by 1292 first-year university students. The graph below shows that the distribution of the data, with a standard deviation of 15.38%, can assumed to be normal. The bell-shaped curve is centered around a mean of 68%. A skewness and kurtosis between -1 and 1 confirm the assumption of normality, which also reflects the norm-referenced character of the test (Dancey & Reidy 2004).

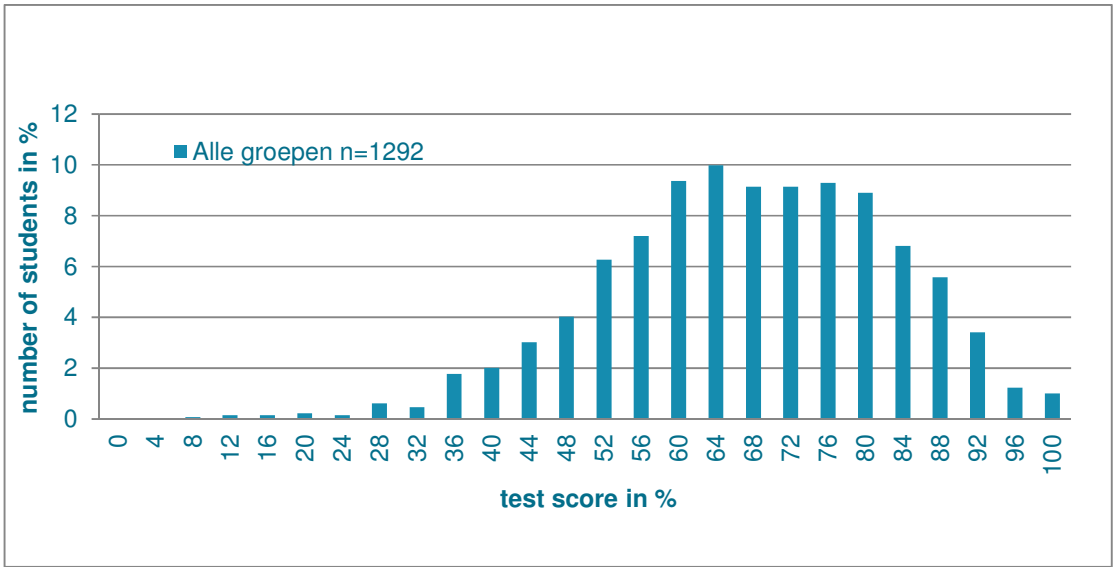


Fig. 2: Test results September 2010

Two studies have then examined the relation between the results of the academic literacy test and the study success of first year university students. On the one hand the faculty of Science has performed regression analyses that have shown that the score on the academic language skills test is, among other factors, a significant predictor of academic achievement. On the other hand, a correlation study for all the faculties that took the academic language skills test indicates a very significant link between the results of the test and students' study success. The outcomes of these investigations will provide meaningful indications on how to improve study success and the efficiency of coaching programs for first year students at the University of Leuven.

This article focuses on the second study. Together with the Teaching and Learning Department of the university of Leuven, the external validity of the academic literacy test has been evaluated. The Teaching and Learning Department has managed to correlate the language test with the results of the January exams (Marx & Huyghe 2011). It appeared that there was a moderately positive but very significant correlation between the language test scores and the exam results ($r=0.37$, $p<0.001$). On average, if a student scored lower on the language test, he scored lower on the exams and vice versa. That way, the claim that an academic language test can be used to select a group of at-risk students can be supported. There were a lot of exceptions to this general tendency however, as can be seen in the scatterplot below.

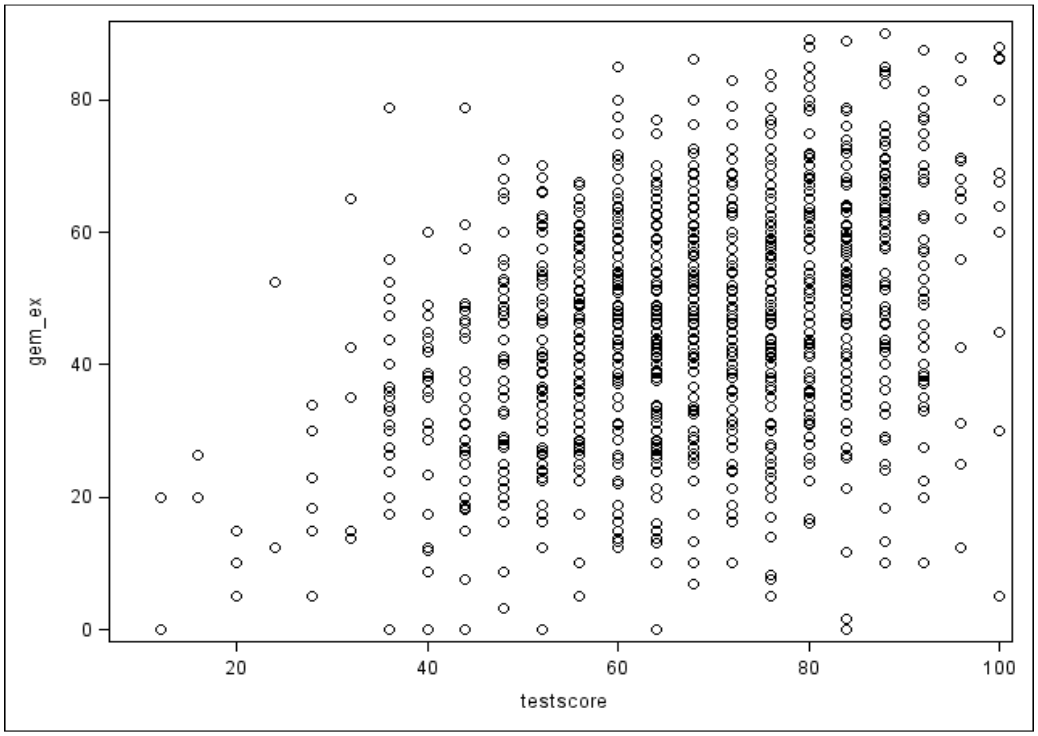


Fig. 3: Scatterplot of test scores and exam results

Important to notice about the graph is that the top left corner is as good as empty. That means that students who fail the academic literacy test do usually not pass their January exams. More specifically 87% of the students that score below 50% on the academic language test, do not pass their January-exams. Research thus confirms the claim that the test can be an early warning signal to those students who have limited academic potential regardless of their field of study. At the same time the group of students that pass the language test have to be aware of the fact that although their language skills are necessary, they are not the only factor of study success. Hence, passing the language test is no guarantee for success: it is a necessary, but not a sufficient condition to pass university exams. This also implies that a very high correlation would be impossible, as study success depends on much more than language skills alone. This is an important message to convey to the different institutions that feel the need to use the academic literacy test, because as Bachman and Palmer (1996) show: "there is often a belief that 'language testers' have some almost magical procedures and formulae for creating the 'best' test" (Bachman & Palmer 1996, 7). The danger in that is that people tend to develop unrealistic expectations, including the test developers themselves.

Consequences and implications: coaching and training

The results of the correlation study mentioned above show that the academic literacy test developed at the Leuven Language Institute may not be a strict predictor of study success but it does select an at-risk target audience. That has implications within the larger educational framework in which the academic literacy test functions. Those students who fail the test are invited to remedy the deficiencies in the field of academic literacy. On the one hand, they can do the exercises on the e-learning platform (ilt.kuleuven.be/taalvast) with learning paths. This e-learning platform keeps individual track of each student. Students are also invited to participate in three interactive language workshops. Both the workshops and the e-learning platform do not only focus on knowledge, but on the various range of strategies that can be used to achieve that knowledge and that enhance the academic potential of the students.

Conclusion

As shown above, developing a valid and reliable proficiency test in the field of academic literacy is quite a challenge. At the same time, the test appears to fulfil a certain need. Since September 2010, over 9000 first year students of the KU Leuven Association have already taken the test at the start of their academic career. Its growing popularity shows that faculties and institutions agree that a tool is needed to meet the diversity of the student influx.The observation that academic language skills are an important indicator of students' future academic achievement, even in exact sciences, suggests that a general academic language skills test could be used to inform both science and non-science students about their chances on early academic achievement.

Even though the test is rather low-stakes and the remedial courses and tools are not obligatory, it is necessary to keep validating the test with several complementary methods. Test development has to be seen as a continuous process of designing, testing and revising in order to meet the intended purpose as accurately as possible.

References

Bachman, L. F. & Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press.

Bonset, H. & de Vries, H. (2009). Talige startcompetenties hoger onderwijs, stichting leerplanontwikkeling (SLO), Enschede.

Brown, J., D. & Hudson, T. (2002). Criterion-referenced language testing. Long, M., H. & Richards, J., C. (Ed.) Cambridge Applied Linguistics, Cambridge: Cambridge University Press.

Brown, J. D. & Hudson, T. (2002). Criterion-referenced language testing. Cambridge: Cambridge University Press.

Dancey, C., P. & Reidy, J. (2004). Statistics without maths for psychology using SPSS for Windows. Third Edition. Essex: Pearson Education.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). Dictionary of language testing. Cambridge: Cambridge university press.

De Wachter, L. (2010). Hoe de K.U. Leuven de 'academische taalvaardigheid' ondersteunt: twee projecten toegelicht. In: Peters E. en Van Houtven T. (red.), Taalbeleid in het hoger onderwijs. De hype voorbij? Leuven: Acco, 153-163.

De Wachter, L. & Cuppens L. (2010). Detectie en remediëring van taalleerzorgproblemen voor eerstejaarsstudenten aan de K.U. Leuven. In: Vanhooren, S. en Mottart, A. (Red.), 24ste conferentie het schoolvak Nederlands. Gent: Academia press, 260-264.

De Wachter, L. & Heeren, J. (2013), Een taaltest als signaal. De ontwikkeling en implementatie van een strategische taalvaardigheidstoets aan de KU Leuven. In: Levende Talen Tijdschrift, 2013/2 (accepted for publication).

Guay, F., Ratelle, C.F. & Chanal, J. (2008) Optimal learning in optimal contexts: the role of self-determination in education. Canadian Psychology 49, 233-240.

Hacquebord, H. & Lenting-Haan, K. (2012). Kunnen we de moeilijkheid van teksten meten? Naar concrete maten voor referentieniveaus. In: Levende talen tijdschrift, 13, nr. 2, 14-23.

Hazenberg, S. & Hulstijn, J.H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. Applied Linguistics, 17, 145-163.

Holder, J.M., Jones, J., Robinson, R.A. & Krass, I. (1999) Academic Literacy Skills and Progression Rates Amongst Pharmacy Students. Higher Education Research & Development 18, 19-30.

Jansen, C. & Lentz, L. (2008). Hoe begrijpelijk is mijn tekst? De opkomst, neergang en terugkeer van de leesbaarheidsformules. Onze Taal, 1, 4-7.

Marx, S. & Huyghe, S. (2011). Eerste analyses van het project TaalVaST. De correlatie tussen taaltest en examenresultaten. Intern rapport DUO. [https://ilt.kuleuven.be/cursus/docs/Eerste_analyses_taalvast_20110902.pdf– consulted: 21/03/13].

Marsh, H. W. & Craven, R. G. (2006). Reciprocal Effects of Self-Concept and Performance From a Multidimensional Perspective Beyond Seductive Pleasure and Unidimensional Perspectives. Perspectives on Psychological Science, 1, 133-163.

McNamara, T. (1996). Measuring second language performance. London, England and New York, NY: Addison Wesley Longman.

McNamara, T. (2000). Language testing. Oxford: Oxford University Press.

Peters E. & Van Houtven T. (2010). Taalbeleid in het hoger onderwijs. De hype voorbij? Leuven: Acco.

Peters, E., Houtven, T. Van & Morabit, Z. (2010). Is meten echt meer weten? Taalvaardigheid van instromende studenten in het hoger onderwijs in kaart gebracht. In: Peters E. en Van Houtven T. (red.), Taalbeleid in het hoger onderwijs. De hype voorbij? Leuven: Acco, 51-66.

Sercu, L., Vyncke, C. & Peters, E. (2003). Testen en evalueren in het vreemdetalenonderwijs. Cahiers voor didactiek, 15. Academisch vormingsinstituut voor leraren (KU Leuven), Mechelen: Wolters Plantyn.

Tyson, W. (2011). Modeling Engineering Degree Attainment Using High School and College Physics and Calculus Coursetaking and Achievement. Journal of Engineering Education, 100, 760-777.

Van den Branden, K. (2010). Taalbeleid in het hoger onderwijs. In: Peters, E. en Van Houtven, T. (Eds.), Taalbeleid in het hoger onderwijs: de hype voorbij? Leuven/Den Haag: Acco, 213-223.

Van Dyk, T. (2010). Konstitutiewe voorwaardes vir die ontwerp van ’n toets van akademiese geletterdheid. Universiteit Stellenbosch. [PhD]

Veenstra, C. P., Dey, E. L. & Herrin, G. D. (2008). Is Modeling of Freshman Engineering Success Different from Modeling of Non-Engineering Success? Journal of Engineering Education, 97, 467-479.

Zhang, G. L., Anderson, T. J., Ohland, M. W. & Thorndyke, B. R. (2004). Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering Education, 93, 313-320.

Katrijn Denies & Rianne Janssen

Katholieke Universiteit Leuven, Leuven, Belgium

katrijn.denies@ppw.kuleuven.be

CEFR Can-Do Statements as a Means of Self-Assessment: is There a Common Understanding, Regardless of the Student's Gender and Educational System?

Bio data

After teaching English and Dutch as a second language for one year, **Katrijn Denies** started working at the Centre for Educational Effectiveness and Evaluation at the KU Leuven to execute the Flemish branch of the European Survey on Language Competences (ESLC). She became the study's national research coordinator in September 2012 and as such presented the study's final results in late May 2013. Meanwhile, she also taught English teaching practice at the KU Leuven. Starting September 2013, she will be working full time on finishing her PhD which will be a thorough investigation of large-scale assessments of students' second language competences.

Rianne Janssen is a professor at the Centre for Educational Effectiveness and Evaluation at the KU Leuven. Her work focuses mostly on educational measurement. She was appointed as the promoter-coordinator of the European Survey on Language Competences in Flanders by the Flemish Ministry of Education. She is also the promoter-coordinator of the Flemish national assessments of the attainment targets, which regularly include assessments of second language skills.

Abstract

The Common European Framework of Reference (CEFR) contains a set of concrete descriptive scales, which, ideally, should allow for unambiguous communication regarding stages in language learning. Yet there is still need for empirical research on how much room for interpretation is left by the descriptors. The aim of this paper is to investigate whether the use of CEFR can-do statements as a means of self-evaluation shows a gender bias and a country or region-related bias, which is commonly found with other measures of academic self-concept.

The study is based on data from the European Survey on Language Competences (ESLC). The ESLC was carried out in spring of 2011 by order of the European Commission. It was the first survey to provide information on students' second language competences that can be compared across fourteen countries, five languages and three skills (reading, listening and writing). As part of the survey's extensive background questionnaire, students were asked to assess their own competences using four can-do statements per skill that were all taken or adapted from the CEFR descriptor scales. The current paper deals with findings resulting from a comparison between the students' measured CEFR-level on the one hand, and their can-do self-evaluation on the other hand. Students were labeled as 'overestimating' their own level, 'underestimating' it or having a 'correct' self-concept for the tested skill.

Multinomial logistic regression analyses on data from over 40000 students revealed that, when they assess their own skills by means of the can-do statements, the students' odds of over- or underestimating their competences correlate significantly with their

gender and educational system (i.e., country or region). In other words, the can-do statements are currently subject to bias: however concrete the statements may already seem, they are still perceived in a way that reflects the students' personal or national norms rather than fixed criteria.

Short paper

Study description

Purpose

The Common European Framework of Reference (CEFR) contains a set of concrete descriptive scales, which, ideally, should allow for unambiguous communication regarding stages in language learning. Yet there is still need for empirical research on how much room for interpretation is left by the descriptors. The aim of this paper is to investigate whether the use of CEFR can-do statements as a means of self-evaluation shows a gender bias and a country or region-related bias, which is commonly found with other measures of academic self-concept (e.g., Marsh, 1998; Chiu & Klassen, 2010).

Data

The study is based on data from the European Survey on Language Competences (ESLC) (European Commission, 2012a; European Commission, 2012b). The ESLC was carried out in spring of 2011 by order of the European Commission. It was the first survey to provide information on students' second language competences that can be compared across fourteen countries, five languages and three skills (reading, listening and writing). As part of the survey's extensive background questionnaire, students were asked to assess their own competences. They did so using four can-do statements per skill that were all taken or adapted from the CEFR descriptor scales.

Method

The current paper deals with findings resulting from a comparison between the students' measured CEFR-level on the one hand, and their can-do self-evaluation on the other hand. This was done for each of the three tested skills: listening, reading and writing. Three out of five languages were explored, namely English, French and German. This is the case because too few countries participated in the ESLC for Spanish (2 countries) and Italian (1 country).

First, both elements in the comparison, i.e. the students' measured CEFR-level on the one hand and their self-estimated CEFR level on the other hand, were determined for each of the three skills. With regard to the former, the ESLC database contains five plausible values per student and per skill but it does not contain one final, estimated CEFR level. Therefore each plausible value was compared to the set standards to achieve a list of five plausible CEFR levels per student and per skill, and from this list, the mode was chosen as the CEFR level to be used in the analyses. Because no tests were aimed at CEFR levels C1 or C2, the highest level was named 'B2 or higher'. The lowest level was named 'pre-A1'.

Additional steps were also needed to determine the students' self-estimated CEFR level. Students were asked to indicate whether or not they felt capable of doing four different tasks per skill by choosing 'yes' or 'no, not yet'. These tasks were formulated as can-do statements which were related to the CEFR descriptor scales and they ranged in difficulty from A1 to B2. The study assumed that for each skill, each individual students' answers would show a Guttman response pattern (Guttman, 1950). Guttman response patterns occur when questionnaire items have a specific order, with respondents who agree to a particular item also agreeing with all lower rank-ordered items. For example, it was expected that students who agreed to the B1-level can-do statement for listening, also agreed to the A2 and A1-level statements. Guttman response patterns turned out to be present in 84%, 89% and 86% of the records for reading, listening and writing respectively. This was deemed enough to apply the following strategy: for each skill, the

number of times that a student answered 'yes' to a can-do statement was taken and this sum was transformed in accordance with Table 1.

N of affirmed can-do statements	Self-estimated CEFR level
0	Pre-A1
1	A1
2	A2
3	B1
4	B2 or higher

Table 1: Conversion of the number of affirmed can-do statements to the CEFR level used in the analyses

Next, records with missing responses were deleted from the dataset. In the sub-datasets per skill, records were removed if students did not assess one or more of the four can-do statements for that particular skill. Also, the records of students who did not indicate their gender in their questionnaire were removed from the dataset. Both steps resulted in a total data loss of about 4%. Table 2 states the final sample size for each of the three tested languages.

Language	N writing	N reading	N listening
English	14103	14595	14462
French	4447	4757	4666
German	6536	6797	6805

Table 2: Final sample size per language

A comparison between the students' measured CEFR level and their self-estimated CEFR level then labeled each student as 'overestimating' their own level, 'underestimating' it or having a 'correct' self-concept for the tested skill.

Results

Multinomial logistic regression analyses were used to reveal whether the odds over being an overestimator or an underestimator rather than having a correct self-concept correlated significantly with the students' gender and educational system. Table 3 shows that this was indeed the case for the given examples, namely for English reading, for German listening and for French writing. For English reading, girls were more likely to underestimate themselves than boys and for French writing, girls were less likely to overestimate themselves than boys. For German listening, on the other hand, the girls' odds of overestimating themselves was larger than that of boys. Furthermore, in several countries and for each example, the students' odds of over- or underestimating themselves differed significantly from the odds for students in the reference country to do so.

Type	Underestimating			Overestimating		
Dependent variable	English reading	French writing	German listening	English reading	French writing	German listening
Female	0.28***	0.16	0.11	0.04	-0.42***	0.15**
Belgium German	1.23***	0.07	/	0.17	-0.74***	/
Belgium French	0.42*	/	-0.16	1.06***	/	0.19
Belgium Dutch	0.93***	0.22	/	-1.34***	-0.49***	/
Bulgaria	0.29	/	0.82***	0.30**	/	0.81
Estonia	0.42**	/	0.29	-0.54***	/	0.21
Greece	0.12	-0.02	/	0.16	0.84***	/
Spain	0.17	0.11	/	0.77***	-0.45***	/
France	0.15	/	/	1.67***	/	/
Croatia	-0.27	/	0.18	0.19*	/	0.68***
Malta	-0.14	/	/	-1.53***	/	/
Netherlands	0.42**	/	0.73***	-0.32***	/	-0.69***
Poland	-0.24	/	-0.81	0.79***	/	0.72***
Portugal	-0.30	REF	/	0.81***	REF	/
Sweden	-0.14	/	/	-1.27***	/	/
Slovenia	REF	/	REF	REF	/	REF

*: p < 0.05, **: p<0.01, ***: p < 0.001, REF: reference category

Table 3: Regression coefficients of Model 1 for the odds of students over- or underestimating their CEFR-level

Type	Overestimating			Underestimating		
Dependent variable	English reading	French writing	German listening	English reading	French writing	German listening
Female	0.28***	0.04	0.08	0.05	-0.25***	0.26***
Skill	-0.12***	0.25***	0.15***	-1.60***	-0.27***	-1.01***
Belgium German	1.27***	-0.70***	/	0.23	0.01	/
Belgium French	0.38*	/	-0.16	0.74***	/	0.20
Belgium Dutch	0.98***	-0.27	/	-0.59***	-0.04	/
Bulgaria	0.28	/	0.86***	-0.42***	/	-0.22
Estonia	0.49***	/	0.33*	-0.16	/	0.06
Greece	0.12	-0.53*	/	0.05	0.83***	/
Spain	0.12	-0.40**	/	0.15	0.16	/
France	0.01	/	/	0.71***	/	/
Croatia	-0.29	/	0.24*	-0.03	/	0.37**
Malta	-0.06	/	/	-0.47***	/	/
Netherlands	0.42**	/	0.67***	-0.11	/	-0.05
Poland	-0.26	/	0.07	0.30*	/	-0.11
Portugal	-0.32	REF	/	0.20	REF	/
Sweden	-0.05	/	/	-0.21	/	/
Slovenia	REF	/	REF	REF	/	REF

*: p < 0.05, **: p<0.01, ***: p < 0.001, REF: reference category

Table 4: Regression coefficients of Model 1 for the odds of students over- or underestimating their CEFR-level

The model presented in Table 3 was estimated without taking into account the students' skill level, however. Adding their level to a second model seemed necessary to deal with the fact that some students may not have had the chance to underestimate or underestimate their level because they performed at level pre-A1 or 'B2 or higher' respectively. If the proportion of such students were random this would not impact much on the analyses, but the ESLC brought to light that there are, in fact, substantial differences between countries with regard to the levels their students attain. Model 2,

which is presented in Table 4, therefore repeated the analyses with the addition of the first plausible value for each student as an indicator of their level for the skill under investigation.

Language	Model 1	Model 2
English reading	0.21	0.53
German listening	0.09	0.27
French writing	0.07	0.28

Table 5: Nagelkerke's pseudo R-Square for Model 1 and Model 2

Table 5 shows that this addition increased the proportion of explained variance. Furthermore, and more importantly, Table 4 shows that significant effects of the students' gender and of their educational system remained present. In some cases, adding the first plausible value to the model resulted in a decrease in the number of countries where the students' odds of over- or underestimating their own skill differed significantly from that of students in the reference country. Overall, however, there still was a substantial effect of 'gender' and 'educational system'. Further research should look into factors that could explain this bias.

Discussion: how applicable are the CEFR can-do statements for accurate self-evaluation?

The ESLC was the first study to enable a large-scale comparison of students' language skills in terms of the CEFR across 14 European countries. It also allowed for the can-do statements that were taken or adapted from the CEFR descriptor scales to be evaluated in different ways. In the current paper, multinomial logistic regression analyses on ESLC-data from over 40000 students revealed that, when they assess their own skills by means of the can-do statements, the students' odds of over- or underestimating their competences correlate significantly with their gender and educational system. This finding persisted after controlling for the students' individual skill level.

The results indicate that the can-do statements are currently subject to bias: however concrete the statements may already seem, they are still perceived in a way that reflects the students' personal or national norms rather than fixed criteria. The main implication of this finding is that caution is required when using CEFR can-do statements as a means for international communication in a context where ambiguity is to be avoided.

References

Chiu, M. M. & Klassen, R. M. (2010). Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries. Learning and Instruction, 20:1, 2-17.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer et al. Measurement and Prediction, IV. New York: Wiley.

European Commission. (2012a). First European Survey on Language Competences: Final Report. Brussel: European Commission.

European Commission. (2012b). First European Survey on Language Competences: Technical Report. Brussel: European Commission.

Marsh, H. (1998). Longitudinal Structural Equation Models of Academic Self-Concept and Achievement: Gender Differences in the Development of Math and English Constructs. American Educational Research Journal, 35:4, 705-738.

Bart Deygers & Koen Van Gorp

Katholieke Universiteit Leuven, Leuven, Belgium

bart.deygers@arts.kuleuven.be

The Influence of the CEFR on Rating Scale Design

Bio data

Bart Deygers has been developing and researching language tests at the Certificate of Dutch as a Foreign Language (CNaVT) since 2009 (CTO, KU Leuven) while coordinating the language policy at Ghent University and co-chairing the CEFR-group within the Association of Language Testers in Europe (ALTE).

Koen Van Gorp has been working at the Centre for Language and Education, Katholieke Universiteit Leuven (KULeuven), since 1991. He specialized in second language learning, teaching and assessment. He received his PhD in Linguistics at the KU Leuven on a study of second language development and knowledge construction (April 2010). Development of teaching materials, teacher training, and project management have been part of his functions at the centre. Currently he is coordinator of preschool, primary and secondary education at the Centre for Language and Education. Since September 2010 he is project leader of the Certificate Dutch as a Foreign Language.

Abstract

The Certificate Dutch as a Foreign Language (CNaVT) offers Dutch task-based language exams for 6 different profiles which have been determined by an extensive needs analysis (Van Avermaet & Gysen, 2006). The task content is co-determined by a pool of subject specialists around the world who verify the authenticity and representativeness of each task and check the items for cultural bias.

For the past years the CNaVT's rating scale has been dichotomous and analytical. Even though this scale has a proven reliability and usability, it was decided to reshape it into a model that would better reconcile the CNaVT's philosophy with its stakeholders' needs: i.e. a clearer alignment with both the CEFR and domain experts' judgements of language performance (Jacoby & McNamara, 1999).

Redesigning the scale has proven to be an extensive undertaking which touches upon all aspects of language testing. Indeed, altering a dichotomous model into a polytomous band rating scale, which merges performance driven exemplifications (Weigle, 2007) with measurement driven descriptors is an operation so all-encompassing that it necessitates rethinking the entire testing process. Simultaneously, working closely with the CEFR has forced the rating scale developers to critically examine the level descriptors so as to operationalize them in a usable rating scale without neglecting known pitfalls such as validity reduction (Lumley, 2002) and a lack of concreteness (Fulcher, 2010).

This presentation focuses on the role of the CEFR in the rating scale redevelopment process, on its strengths, but also on its shortcomings which prevent it from being a readymade assessment tool. The presentation will include data resulting from the development and validation process. This includes focus groups with subject specialists, stimulated recall interviews with raters as well as qualitative test analyses (i.e. inter and intra rater reliability, correlation coefficients etc).

Short paper

Rating scale typologies

Rating scales can be classified according to different parameters, such as the way in which the scoring criteria have been established or the way these criteria are presented to the rater. Naturally, these different types can be combined and modified to match the idiosyncrasies of each individual test.

Measurement driven rating scales have been drawn up by language experts and are typically not derived from real-life performances, which is the very basis of performance driven scales (Fulcher, Davidson and Kemp 2010, Weigle 2007). Since measurement driven scales are founded in theory and abstraction, their level descriptors may be too distant from reality. Conversely, given that performance driven scales are based on actual performances, their descriptions might be too detailed to allow for generalization (Fulcher et al. 2010).

Holistic rating scales compel raters to judge a performance as a whole, whereas their analytic counterparts take into account separate features of language, such as grammar, vocabulary and structure (Alderson, Clapham and Wall 1995). Previous studies have shown that the analytic scales are often more reliable than holistic ones, offer richer L2 diagnostic information and are better suited for novice raters (Barkaoui and Knouzi 2011, Barkaoui 2010, Knoch 2009, Weigle 2002). Holistic scales on the other hand, perform better than analytic ones in terms of authenticity and rating speed (Knoch 2009, Weigle 2002). A third possible way to categorize rating scales is according to the number of scoring categories they employ. "Items that are scored in two categories - right or wrong - are referred to as dichotomously scored items. Items that are scored in multiple-ordered categories are referred to as polytomously scored items" (Tang 1996: 2).

Whether or not a rating scale is performance driven or measurement driven, holistic or analytic, dichotomous or polytomous, it is always the rater and not the rating scale who decides on the score (Fulcher at al. 2010, Lumley 2002). Naturally, the quality of the descriptions, their level of complexity and abstraction will influence the consistency and accuracy of the rater (Alderson et al 1995, Fulcher et al 2010). Additionally rater training has proven to be of great value when streamlining the interpretations of rating criteria (Lumley 2002, Shohamy, Gordon and Kraemer 1992, Weigle 1994). Without rater training, it would be up to each individual rater to decide on the meaning of frequently used but unquantifiable terms such as "adequate", "good" and "sufficient". Even with such a training it is difficult to overcome the problems associated with vagueness in descriptors.

An Asymmetrical Framework

Upon its publication, the CEFR was to be a document that addresses concerns about multilingualism, stimulates the use of a common metalanguage, helps curriculum development and promotes professional mobility within Europe (Little 2007, Fucher 2004, Milanovic 2001). More than a decade later its actual use differs from these original intentions. As more and more schools, test developers and policy makers use the CEFR it is regarded as more than the theoretical model it actually is (Fulcher 2004) and has become a fixed standard in European language education and language testing. Still, the CEFR, being a measurement-driven language-independent model of L2 acquisition, lacks the empirical foundation and descriptional specificity to act as a real framework (Alderson 2007, Little 2007), let alone a scoring tool (Papageorgiou 2010, Weir 2005).

For one thing, the relative distance between the different CEFR levels is inconsistent (Fulcher 2004). This causes fundamental problems in a rating context since polytomous IRT analysis generally assumes that the distribution between different scoring levels is equal (Huyn 1994 & 1996, Tang 1996) or at least known (Roberts, Donoghue & Laughlin 2010).

Furthermore, the level descriptors often show overlaps and gaps (Alderson 2004), both of which may create the vagueness a rating scale constructor whishes to avoid.

"When the scales, in particular, were examined closely, it became apparent that many terms lacked definitions, there were overlaps, ambiguities, and inconsistencies in the use of terminology, as well as important gaps in the CEFR scales."
(Alderson 2007: 661)

Finally, the CEFR is asymmetrical in the attention it gives to receptive and productive skills. It focuses heavily on production while the receptive skills remain underdeveloped (Weir 2005, Alderson 2004, Staehr 2008, Milton 2010). The CEFR lacks usable specifications for quite a few skills that may be operationalized in receptive tasks, i.e. text complexity (Alderson 2004, Weir 2005, Alderson 2006, Davidson & Fulcher 2007), lexis (Alderson 2007, Milton 2010) and subject matter (Weir 2005, Fulcher 2004).

CNaVT Rating scale construction

The Certificate of Dutch as a Foreign Language (CNaVT) offers five functional task-based language tests (Van Gorp & Deygers 2013) that operate according to Bachman and Palmer's (2010) can-do typology. These tests correspond to five profiles and fall into three categories: societal, professional, and academic language use. The profiles have been determined by a needs analysis among end users (Van Avermaet & Gysen 2006), which continues to shape the exams to date. Currently, the CNaVT is a pass/fail exam: candidates either pass the examination in the domain of their choosing or they do not.

In 2009, the subject specialists of the two academic profiles suggested altering the binary approach of the existing dichotomous analytic rating scale so it would align more closely with their "indigenous criteria" (Jacoby and McNamara 1999). Around the same time quite a few stakeholders voiced their wish for the different tests to be more explicitly linked to the CEFR (a trend also observed by Fulcher 2004). More recently, the government organisation funding the exams has decreed that over the coming years the pass/fail approach should be abandoned in favour of a system in which each test contains two cut scores, each one linked to a CEFR level. These developments instigated both a revision of the testing process and a reconceptualization of the rating scale (see Deygers, Van Gorp, Luyten and Joos 2013 for a full discussion of the rating scale construction and validation process). The new rating scales were to reconcile the subject specialists' criteria with both the stakeholders' wish for a clearer CEFR alignment and with the test sponsor's demand for a double cut off score at two CEFR levels per test. Even though all rating scales are in the process of revision, this paper solely focuses on the scale of the new Dutch for academic purposes (DAP) test.

The composition of the DAP's team of raters may change from one year to the next. Since the judgment of novice raters is more reliable when using an analytic rather than a holistic scale (Barkaoui 2010), the new scales are analytic in nature. The criteria for these scales are derived from focus groups with subject specialists (N = 13), subject specialist questionnaires (N = 178) and literature reviews (Deygers et al. 2013). Each criterion can be scored on four levels, the third being up to the minimum standard, the fourth being above and the first and second below. Each scoring category corresponds to a CEFR level. In the case of the DAP test, three corresponds with the B2 level of the CEFR, four with C1.

After an iterative development process, the DAP rating scale was piloted using 4 trained raters who rated 250 tasks using both the original dichotomous scale and the newly developed polytomous scale. In order to avoid sequence or contamination effects, two raters first used the polytomous scale while the other two started with the dichotomous scale. Irrespective of the order in which the scales were used, the dichotomous scale consistently showed to be more reliable and to yield higher inter-rater agreement (Deygers et al. 2013).

Following the rating process, the four raters took part in a focus group. They preferred the dichotomous scale when judging written performances but the polytomous one for speaking tasks. All raters preferred the polytomous approach in theory because it allows for a more fine-grained judgment. In practice, they all reported confusion when using the CEFR-based level descriptors.

A second and third trial followed the initial pilot of the rating scale. Each new trial focused on rewriting the level descriptors so they would become more easily interpretable by novice raters. Before each trial, the raters received an intensive two-day rater training during which they reported vagueness in the level descriptors and suggested ways to reformulate the descriptions. Often these suggestions meant clarifying the difference between one level and the next, providing concrete examples and adding language-specific expectations. In the second trial, two trained raters judged 76 spoken performances and in the third trial two trained novice raters judged 27 written argumentative tasks and 28 presentation tasks.

After each trial the raters now reported to prefer the polytomous scale over the dichotomous one. They did not report feeling uncertain or confused when using the adapted level descriptors. Nonetheless, quantitative analysis of the rating process shows that the descriptors of grammar and vocabulary caused problems. For grammar, the distinction between level 2 (B1) and 3 (B2) was considered too harsh. For vocabulary, all descriptors remained too vague. Other CEFR tables such as "Orthographic control" and "Coherence and cohesion" also appeared quite challenging indeed to operationalize.

Discussion: The use of the CEFR for rating scale design

Even though the CEFR "was not designed specifically for test specifications and language testing contexts" (North 2004 in Papageorgiou 2010: 273), there is an apparent need within Europe among stakeholders to demand a clear link between a test score and a CEFR level.

"For many producers of tests, one of the dangers lies in the desire to claim a link between scores on their tests and what those scores mean in terms of CEF levels, simply to get recognition within Europe. They do not have any choice in this, for if institutions begin to believe that the CEF is the truth against which all else must be measured, failure to claim a link to the CEF would equate to a commercial withdrawal from continental Europe." (Fulcher 2004: 260)

In the case of the CNaVT, the endeavor to link the test with the CEFR has surpassed the "intuitive guess" Fulcher (2004) observes. Each CNaVT test has undergone an extensive standard setting process and the rating scales combine input from subject specialists, language specialists, raters and the CEFR level descriptors. By working closely with the CEFR, the developers of the rating scale have critically examined the its level descriptors in order to operationalize them in a usable rating scale while avoiding validity reduction (Lumley, 2002) lack of concreteness (Fulcher, 2010) and other known pitfalls of rating scale construction.

The major shortcoming of the CEFR when used as a source for rating scale development appears to be its unsound theoretical foundation. It is partly based on empirical findings but at its core are the intuitions of language experts (Alderson 2004, Fulcher 2004, Hulstijn 2007, Little 2007, North 2007). This leads to inconsistency and vagueness on a meta and micro level. On a meta level, the unequal distance between levels causes problems for a polytomous IRT analysis. On a micro level, not all level descriptors can readily be operationalized in a rating scale.

One example of this is the CEFR's description of grammatical accuracy. The difference between "relatively high degree of grammatical control [without] mistakes which lead to misunderstanding" (lower end B2) and "generally good control […] errors occur, but it is

clear what he/she is trying to express" (higher end B1) is too tentative to use in a rating context. Using either the lower B1 and the upper B2 or the upper B1 and upper B2 prove equally problematic. All raters involved in the pilot study claimed that the difference between 2 (B1) and 3 (B2) was either too vague or too harsh to be usable. Even though the criterion "grammar" caused some correlational problems among the raters, "vocabulary" yielded the lowest inter-rater agreement of all criteria. Indeed, the CEFR "provides little assistance in identifying the breadth and depth of productive or receptive lexis" (Weir 2005: 292).

Conclusion: a common basis?

The CEFR is a theory on second language acquistion, partly based on empirical data, partly on theoretical conceptions and partly on intuition (Hulstijn 2007, Little 2007, North 2007). It takes on a positive and descriptive approach to language learning by focusing on what learners can do at a given level. This has forced language teachers not to only think of their students in terms of deficit, but also in terms of accomplishment. Throughout Europe language practitioners and policy makers now not only know that the CEFR exists and use its terminology, they may also see what it entails and might even wish for classroom and testing practice to adhere to its logic. And that is where the problem begins.

For one thing, no theoretical model can strive towards universality without trading in specificity for generic applicability. Because of this, every CEFR descriptor used in the CNaVT rating scale development was too underdefined to be used without adaptation. For each criterion language-specific additions had to be made, differences between levels had to be clarified and examples had to be provided. Only then were raters able to maintain an acceptable level of consistency.

Furthermore, the CEFR occupies a somewhat dubious position in terms of malleability. In the minds of many stakeholders and policy makers the CEFR-levels appear etched in stone, B2 occupying an especially elevated position. At the same time however there is general agreement that the broadness of CEFR level descriptors allows for multiple interpretations, forcing users into interpretation and specification. And when generally accepted levels are universally interpreted differently, the CEFR can only provide "a common basis" (Milanovic, 2001: 1) on paper.

References

Alderson, C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). The Development of Specifications for Item Development and Classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment Reading and Listening. Final report of the Dutch CEF construct project. Unpublished Document. Retrieved from http://eprints.lancs.ac.uk/44/1/final_report.pdf

Alderson, J. C. (2007). The CEFR and the Need for More Research. The Modern Language Journal, 91(4), 659–663.

Alderson, J. C., Clapham, C. & Wall, D. (1995). Language Test Construction and Evaluation. Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. Language Assessment Quarterly: An International Journal, 3(1), 3–30.

Bachman, L. & Palmer, A. (2010). Language Assessment in Practice. Oxford University Press, USA.

Barkaoui, K. & Knouzi, I. (2011). Rating scales as frameworks for assessing L2 writing: examining their impact on rater performance. Presented at the ALTE 4th International Conference, Kraków, Poland.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. Language Testing, 27(4), 515–535.

Davidson, F. & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. Language Teaching, 40(3), 231. doi:10.1017/S0261444807004351

Deygers, B., Van Gorp, K., Luyten, L. & Joos, S. (2013). Rating scale design: a comparative study of two analytic rating scales in a task-based test. In E. Galaczi & C. Weir (Ed.), Exploring Language Frameworks. Proceedings from the ALTE Kraków Conference, July 2011. (Vol. 36, pp. 273–289). Cambridge: Cambridge University Press.

Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. Language Assessment Quarterly, 1(4), 253–266.

Fulcher, G., Davidson, F. & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. Language Testing, 28(1), 5–29.

Hulstijn, J. H. (2007). The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency1. The Modern Language Journal, 91(4), 663–667.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. Psychometrika, 59(1), 111–119.

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. Psychometrika, 61(1), 31–39.

Jacoby, S. & McNamara, T. (1999). Locating Competence. English for Specific Purposes, 18(3), 213–241.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. Language Testing, 26(2), 275–304.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. The Modern Language Journal, 91(4), 645–655.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? Language Testing, 19(3), 246–276.

McNamara, T. & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. Language Testing.

Milanovic, M. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In G. Pallotti (Ed.), Communicative proficiency and linguistic development: intersections between SLA and language testing research. Rome: Creative Commons.

North, B. (2007). The CEFR Illustrative Descriptor Scales. The Modern Language Journal, 91(4), 656–659.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. Language Testing, 27(2), 261–282.

Roberts, J. S., Donoghue, J. R. & Laughlin, J. E. (2000). A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses. Applied Psychological Measurement, 24(1), 3–32.

Shohamy, E., Gordon, C. M. & Kraemer, R. (1992). The Effect of Raters' Background and Training on the Reliability of Direct Writing Tests. The Modern Language Journal, 76(1), 27–33. doi:10.2307/329895

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. Language Learning Journal, 36(2), 139–152.

Stæhr, L. S. (2009). Vocabulary Knowledge and Advanced Listening Comprehension in English as a Foreign Language. Studies in Second Language Acquisition, 31(4), 577–607.

Tang, L. K. (1996). Polytomous Item Response Theory (IRT) Models and their applications in large-scale testing programs: review of literature. New Jersey: Educational Testing Service.

Van Avermaet, P. & Gysen, S. (2006). From needs to tasks: Language learning needs in a task-based approach. In K. van den Branden (Ed.), Task-Based Language Education: From Theory to Practice. Cambridge: Cambridge University Press.

Van Gorp, K., Deygers, B. & Kunnan, A. (2013). Task Based Language Assessment. In The Companion to Language Assessment. New Jersey: Wiley-Blackwell.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. Language Testing, 11(2), 197–223.

Weigle, S. C. (2002). Assessing Writing. Cambridge University Press.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing, 22(3), 281–300.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. Language Testing, 27(1), 119–140.

Dan Frost & Jean O'Donnell

Université de Savoie, Chambéry, France

dan.frost@univ-savoie.fr - jean.o-donnell@univ-savoie.fr

Combatting the "Can't do Mentality": Expert, Peer & Self-Assessment in a French University Context

The "ELLO" Project (étude longitudinale sur la langue orale)

Bio data

Dan Frost has a French doctorate in English for Specific Purposes and Teaching Theory. After reading Languages and Linguistics at York University, he taught English in Thailand and Sweden before settling in France, where he obtained a Masters in English Phonetics at the University of Provence. He worked as a secondary school teacher for two years in Le Havre, then as a professeur agrégé for ten years in the Computer Science Department at the IUT2 in Grenoble and he is currently a senior lecturer (maître de conférences) in the department of Applied Foreign Languages (LEA) at the Université de Savoie, Chambéry, France. He also works with trainee teachers, especially in the area of teaching pronunciation. His main research interests are teaching pronunciation, oral English and computer-mediated learning and motivation.

Jean O'Donnell studied in St Patrick's College, Maynooth, Ireland from where she graduated with a primary degree in mathematics and languages (French/Irish) and the national Irish teaching qualification for second-level education. This was followed by a period in France as an assistante, maître de langue and ATER. She obtained a degree in English, a Masters in Applied Languages, the CAPES in English and a French doctorate in Applied Linguistics from Université Stendhal, Grenoble 3. She is now a senior lecturer (maître de conférences) in the Department of Applied Foreign Languages (LEA) at the Université de Savoie, Chambéry, France. Her main research interests are language testing and computer-mediated language teaching and learning.

Abstract

The CEFR was adopted by the French National Ministry of Education and Research (MENR) in 2005. Its relevance to the French context (Goullier 2007) and its applicability (Taylor 2011, Petit 2007, Luoma 2004) have been discussed by several authors. We will analyse the practical aspects of using the CEFR as a tool in a longitudinal cohort research project involving the assessment of English spoken production among Applied Foreign Languages undergraduates. As practitioners, we were struck by the lack of accurate information, specifically in relation to spoken production, to describe the levels of our students before, during and at the end of their three-year degree. Obtaining a language profile for our students at regular intervals would enable us to design and evaluate our courses more effectively. Consequently, as researchers we started a study in 2011 which involves collecting and assessing several samples of spoken data for each student from first to third year. Three types of assessment are currently being carried out: "expert assessment", "peer assessment" and "self-assessment".

This paper explains the reasons for choosing the CEFR as our basic measurement tool after having considered other possibilities. We discuss the advantages and disadvantages encountered so far according to the type of evaluation that is being carried out. For

example does experts' use of the scale coincide? How do peers perceive and apply the scale? Are self-assessments using the scale reliable? What features of the scale are experts satisfied/comfortable with? Finally there were attempts to make constructive suggestions in the light of our preliminary findings.

Short paper

Description of research
As researchers, we have set out to create an accurate linguistic portrait of our students upon arrival in first year, during second year and prior to departure in third year. This means collecting and analysing three consecutive cohorts of students over three academic years - a project requiring 5 years for data collection alone. We will thus be able to confirm or refute our impressions related to competence levels, track our students' progress and tailor our programmes to their needs. This project was launched in 2011-12 and is now in year two of data collection.

In order to obtain speaking samples from the participants, two different activities were selected. The "monologue" consists of describing a short video with a storyline and the "interaction" in pairs consists of a conversation on a topic of interest to this age group. Both tasks proved to be successful in eliciting oral production in the WebCEF [1] research program.

Students were filmed using webcams and recorded using digital microphones in their usual classroom setting with their regular teachers. All recordings were uploaded to Moodle at the end of each session. Complementary information, to be used at a later date, was obtained for each student and includes online language-motivation and language-profile questionnaires, Oxford Placement Test listening test, Cambridge FCE listening test and the Dialang placement, listening and vocabulary tests. During a two-hour session, students were familiarised with the CEFR scales and practised applying them. During a further two-hour session each student evaluated a series of recordings using the scales : a) his/her monologue b) his/her interaction c) his/her partner's interaction d) 5 monologues representative of the first 5 levels of the CEFR e) 5 interactions representative of the first 5 levels of the CEFR. The data obtained from these evaluations and from those of the experts was subjected to a preliminary analysis.

Since the methodology we have chosen to adopt is a longitudinal study, our research questions are not hypotheses which we are aiming to prove or disprove with a randomised control study. This longitudinal cohort study will provide quantitative and qualitative data which will, by the end of the study, help us to answer the following research questions:

1. What practical features of the scale are we satisfied with regarding our needs and students' needs when it comes to oral production?
2. How do the results of the self-assessments, peer assessments and expert assessments when using the scale compare?
3. Does the application of the CEFR scales in a process of self, peer and expert assessment lead to these subjects noticing their own strengths and weaknesses (and thus to improvements in their performance)?
4. How will this process affect their motivation?
5. How will the results of this study compare to institutional exam grades?

Discussion
The title of this paper may seem provocative, but we must stress that it refers to the domain of assessment in France and not to language teaching as a whole. The

[1] http://www.webcef.eu/

assessment system which is almost exclusively used throughout secondary and tertiary education in France is based on a mark out of 20. Students are typically required to obtain la moyenne (the average) in order to pass a module, to successfully complete a year or even to graduate at university level. The word moyenne clearly does not correspond to the statistical term "mean", but more to the notion or a threshold or pass-mark. A mark 20/20 is extremely rare and marks are deducted for mistakes, omissions, etc. so assessment is usually based more on what a student doesn't know or can't do. This approach is obviously very different from the principles underpinning the CEFR and the ELLO project involves students assessing their own performance and that of their peers using the descriptors and scales laid out in the CEFR documents, i.e. having to adopt a "can do" and not a "can't do" mentality.

Assessing oral production is a necessary challenge for educators. Oral skills have always lagged behind reading and writing in France, partly because France is a country which has traditionally protected its language in the public domain and laws such as "La Loi Toubon" have led to a strict limitation of foreign languages in the media, the dubbing of television and cinema, etc. This situation is at last beginning to change with online informal learning of English beginning to make an impact on French students' use of English (Sockett 2011). The French Ministry of Education and Research has officially adopted the CEFR levels and a ministerial decree (MENR 2005) suggests the level that a student should have in his/her first and second foreign language at key stages:

- at the end of elementary school, A1 in the foreign language studied;
- at the end of compulsory schooling, B1 in the first foreign language studied, A2 in the second foreign language studied;
- at the end of secondary education, B2 for the first foreign language studied and B1 for the second foreign language studied.

(Goullier 2007: 38).

A recent survey on English pronunciation teaching practices in 31 European countries (Henderson et al. forthcoming) found that the vast majority of teachers who answered used no officially agreed scale to assess pronunciation, but of the 30% of teachers in France who did report using a scale, they all mentioned the CEFR. However across the education system in France as a whole, the CEFR is not widely used in institutional documents and the 20-point system prevails.

A possible consequence of the above situation is that as practitioners in the French university system we are faced with a frustrating lack of data about the language proficiency of our Applied Foreign Language (LEA) students before and upon completion of their three-year language degree. We have little or no precise indication in particular as to their competence in oral production, a crucial component of their programme. At the beginning of the academic year we had become increasingly convinced of a discrepancy between the level of our incoming students and that set out by the MENR, particularly regarding speaking. A considerable number of our students upon their arrival in first year have difficulty expressing themselves orally even in the simplest of terms.

It has frequently been stated by experts in the field of assessment that speaking is one of the most complicated language skills to test (Lado 1961, Shohamy, Reves & Bejarano 1986; Alderson & Bachman 2004). The major difficulties put forward are usually related to the fleeting and multidimensional nature of speech, the distinction between performance and competence, the non-verbal aspects, the cost, the material and time required, the training of testers, the choice of a suitable task and of course a valid and reliable grading or rating system. We would argue that some of the factors which have rendered the task so difficult in the past are no longer as relevant.

Nowadays, capturing the fleeting nature of speech and its non-verbal aspects can be overcome by the use of computers and webcams. In addition, as the current generation

of students is familiar with social networks, smartphones and webcams, there is considerably less psychological stress generated when filming and recording themselves than in the past (Develotte et al 2010). The question of cost must still be overcome but is not insurmountable. Webcams and good quality microphones are affordable and most language teachers have access to multimedia labs, at least in Europe. The issue of choosing a suitable task is also less problematic than in the past. Over the years many interesting projects (WebCEF in particular) have selected tasks which have proven to be suitable. As for rating, recent studies have shown that well-organized rater-training can improve reliability (O'Sullivan 2012).

We were optimistic that involvement in this project, particularly the rating process, would be a rich pedagogical experience for the participants and lead to increased motivation. After considering several rating scales we decided to use the CEFR for several reasons. It has been officially adopted by the MENR. We were familiar with its use thanks to the WebCEF project. In theory it is one of the best-known frameworks among students and colleagues and so we were eager to determine how practical it was to use in a specific university setting in testing oral performances. It would allow us to share our results and encourage colleagues to refer to the CEFR more frequently when designing programs, setting objectives and discussing results. More importantly, the CEFR's positive "can do" descriptors are in keeping with the principles of the ELLO project and its desired outcome.

Using the CEFR scales requires practice and discussion. Initial sessions among colleagues were organized to link our tests with the CEFR scales. Particular attention was paid to "can do" statements related to the qualitative aspect of oral production - range, accuracy, fluency and coherence. Holistic descriptors proved useful when doubts arouse and the illustrative scales for information exchange were useful when selecting criteria to evaluate the interaction. A considerable advantage was the presence of an experienced participant of the WebCEF project from whom we could obtain advice to ensure reliability during our initial collective rating sessions.

Throughout this preliminary phase a number of strengths and weaknesses of the CEFR became obvious when adapting it to our needs. The expert ratings of the students' speaking activities using the scales did not prove difficult. They enabled us to clearly confirm our impression as to the overall level of our incoming students. For the 2011 cohort of students upon arrival, almost 90% are below B2 and centered almost equally around A2 and B1 in relation to the monologue and the interaction. In both 2011 and 2012 the results of the Dialang listening comprehension test were comparable to the monologues and interactions. However the results of the Dialang vocabulary test revealed significant differences with over 60% in the B2 + C1 range. The evaluation of the 2012 interactions is not yet complete as 2012 saw a considerable increase in student numbers. The monologues of the 2012 cohort of students upon arrival, show that almost 70% are below B2. Once again the levels of the students are centered around A2 and B1 with this time considerably more students in A2 than in B1. More granularity is required within the scales to obtain a precise portrait at these lower levels. Adapting a branching approach is an option but we found that this was difficult to apply because establishing cut-off points within potential micro-levels as the authors of the CEFR point out, is a subjective business and would require further validation. As experts we used the sign "+" but we thought it would be too difficult to expect the participants to apply it accurately.

Using the wide range of "can do" statements ensured the face and construct validity of the oral tasks, however the subjectivity or vagueness of certain terms ("quite possible", "generally", etc.) can lead to different interpretations. During the collective expert rating sessions, our WebCEF expert was stricter than we were on several occasions. Also the sheer quantity of descriptors can at times lead to a sense of being overwhelmed and of losing focus while rating. Although there are "can do" statements related to "phonological control" (CEFR: 117) for example, pronunciation features tend to go unnoticed in the

general grids and we feel that more weight should have been given to intelligibility and phonological accuracy.
No performances received a C2 rating from the experts although we felt this could have been the case for one or two participants. The descriptors at the upper end of the scale for oral production cannot be held totally responsible; it is rather a question of selecting speaking activities that will allow students to perform across the entire spectrum from A1 to C2.

The self-assessments of the monologues and interactions revealed results comparable to those obtained by the experts in terms of creating an accurate general portrait of the class. Upon further inspection however, students were not precise when rating a given individual performance. On each occasion less than half the students' ratings corresponded exactly to those of the experts. The students seem to perceive the overall level of the class but lack accuracy. Among those who were inaccurate there was a slight tendency for the weaker students to overrate performances and for the better to underrate. When it came to students' ratings of partners' interactions, these proved to be the most generous, with a fifth in the B2 band. This did not prevent over three-quarters being rated below B2.

The peer assessments involved the cohort of first-year students assessing a selection of five monologues and five interactions. Unknown to them, each represented a CEFR level (excluding C2 as we had no such sample). Overall, slightly more than half the students gave the same rating as the experts to the A2 and B2 monologue and to the A2 interactions. Worth noting was that almost 90% overrated the few A1 monologue performances, several C1 performances were rated C2 – particularly for the interactions and the most frequent level attributed was A2 followed by B1. We examined whether the lower level participants overrated their peers and the higher level participants underrated their peers, but no distinct pattern emerged. It would seem once again that students have an overall impression of the level of group and perhaps feel safer when giving an A2/B1 rating.

Conclusion
Having used the CEFR as the bases for elaborating exams for an extended period we would argue that it is not time for an entirely new framework; the CEFR already exists and there are many valid practical and pedagogical reasons for using it. Though much remains to be done over the next four years, our study so far has shown that inter-rater reliability in testing inspired by the CEFR is a feasible objective among experts provided there is minimal training. Nevertheless it would seem that for students to achieve greater accuracy in rating using the CEFR, they will require further practice. Reliability although desirable, was not the sole object of involving the students in the evaluation process. Becoming aware of their characteristics as language learners and noticing (Guichon & Cohen 2012) are potential positive side effects. Several students felt the recorded performance did not reflect their true competence at the time of the activity. Nevertheless a questionnaire revealed that understanding and applying the "can do" statements was perceived as a useful although in part, unpleasant experience - a "wake-up call" - judging by the number of self-incriminating remarks and resolutions to improve.

Further development of technology to produce more user-friendly computerized Dialang-style descriptors would simplify the task for raters when evaluating speaking. It would mean less wading through pages of descriptors. Further promotion of the CEFR within the professional world would enable employers to have a reliable yardstick to grasp the true language proficiency of candidates and would reassure candidates that they are being judged fairly. More importantly in a French context, further training of teachers in the use of the CEFR and raised awareness among language students related to what they can do rather than what they cannot, would enhance the entire evaluation process to make it a more positive experience for all involved.

References

Coombe, C., Davidson, P., O'Sullivan, B. & Stoynoff, S. (Ed.). (2012). The Cambridge Guide to Second Language Assessment. Cambridge: CUP.

Council of Europe. (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

Goullier, F. (2007). Le Cadre européen commun de référence pour les langues, instrument de normalisation ou document instrumentalisé pour une normalisation de l'enseignement et de l'évaluation?, Cahiers de l'APLIUT, 26(2), 12-22.

Guichon, N. & Cohen, C. (2012). Enhancing L2 learners' noticing skills through self-confrontation with their own oral production performance. Cahiers de l'APLIUT, 31(3), 87-104.

Develotte, C., Guichon, N. & Vincenta, C. (2010). The use of the webcam for teaching a foreign language in a desktop videoconferencing environment. ReCALL 22(03), 293-312.

Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunningham, U. & Curnick, L. (forthcoming). English Pronunciation Teaching in Europe Survey: Inside and Outside the Classroom.

Lado, R. (1961). Language Testing: The construction and use of foreign language tests. London: Longman.

Louma, S. (2004). Assessing speaking. Cambridge: CUP.

MENR. (2005). Décret n° 2005-1011 du 22 août 2005 relatif à l'organisation de l'enseignement des langues vivantes étrangères dans l'enseignement scolaire, à la réglementation applicable à certains diplômes nationaux et à la commission académique sur l'enseignement des langues vivantes étrangères. Bulletin officiel de l'Éducation nationale. N°31, 1 septembre 2005. Retrieved March 18, 2013, from http://www.education.gouv.fr/bo/2005/31/MENE0501621D.htm

Sockett, G. (2011). From the cultural hegemony of English to online informal learning: Cluster frequency as an indicator of relevance in authentic documents, ASp, 60, 5-20.

Taylor, L. (2011). Introduction, in Taylor, L. & Weir, C.J. (Ed.), Examining Speaking: Research and practice in assessing second language speaking. Studies in Language Testing 30. Cambridge: UCLES/CUP, 1-35.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral interview, Studies in Second Language Acquisition, 10(2), 149-164.

Petit, M. (2007). La correction linguistique dans le Cadre européen commun: quelle conception, quels critères ?, Cahiers de l'APLIUT, 26(2), 62-80.

Shohamy, E., Reves, T. & Bejarano,Y. (1986). Introducing a new comprehensive test of oral proficiency. English Language Tea

Daniela Forapani

Università di Parma, Parma, Italy

daniela.forapani@unipr.it

Designing an Online Italian L2/LS Placement Test in Line with the CEFR Standards. Suggestions for Monitoring Reliability and Ensuring its Validity in the Perspective of an International Application.

Abstract

The placement tests are conceived to measure the knowledge of language that students have, in order to enter them into the corresponding course and guarantee a standard level of ability within the class. These tests might be based on the syllabus adopted by the host institution or on unrelated material.

Given that the University of Parma language courses are structured in line with the CEFR, the online test developed by the Language Centre relates to the L2 Italian Syllabus (Lo Duca, 2009), which divides linguistic content (tasks and texts, linguistic functions, form, structure and meaning) into the six levels A1-C2. This paper sets out the guidelines that item writers/editors must refer to while designing the test, including: a) a list of the technical specifications (number of sections required in the test, number of items per section, types of abilities that have to be tested, features of the linguistic elements that must be included), b) a checklist to ensure that items and CEFR levels correspond and c) various indications to assess the reliability of the test and ensure its overall validity. As the online placement developed in Parma has been used both by the University of Zagreb (Croatia) and the University Falun (Sweden), the international impact of this model will also be investigated.

Short paper

I researched into placement tests at the Language Centre of the University of Parma, Italy, because I had to solve a problem.

The Italian placement test we used to use, to divide Erasmus into classes, plunged into a crisis when the percentage of Spanish-speaking exceeded 50% of the total.
As it is well known, Italian and Spanish are cognate languages, therefore these students, on the basis of their passive knowledge and linguistic affinity, easily achieved a B1 level (yet without knowing how to use the present indicative of the auxiliary verbs when speaking or writing).

The classes formed on the basis of this type of test were obviously problematic.
They were problematic for the teachers, who had to manage them, and for German/Slavic students who found themselves taking a course with other fellows who didn't have the required skill level in grammar and writing.

In terms of content progression, the predominance of Spaniards pushed the teachers to accelerate the rhythm of teaching, yet still not finding a way to remedy the basic gaps in these students' knowledge.
Since the initial variation in levels within the class conditioned the learning process, it was necessary to re-consider the structure of the whole placement, which was a non-CEFR based one.

The steps undertaken to design a new CEFR-based test were aimed to form more homogenous classes.

In order to tag every input and linguistic skill to the appropriate level, we consulted the following CEFR-based sillabi:

1. the L2 Italian syllabus for foreign students by Lo Duca (Lo Duca, 2006)
2. the CILS Certification syllabus by the University for Foreigners of Siena (2009),
3. the PLIDA Certification by the Società Dante Alighieri (2004).

Comparing these three syllabi, we noticed that some items were placed in a different level by every syllabus.
To give an example, Table 1 proposes the tagging of the item related to the imperfect tense, which has been placed in a different level by every syllabus (A1, A2 and B1):

Syllabus	**Imperfect**	**CEFR level**
Lo Duca	Uses and functions of the imperfect indicative: to describe the past.	A1
CILS	Imperfect indicative.	A2
PLIDA	Imperfect indicative.	B1

Table 1 - Tagging of imperfect tenses in Italian CEFR-based syllabi

As we couldn't classify the imperfect tense item on the basis of the sillabi, we consulted three up to date Italian manuals for foreign learners (Chiaro!, Domani e Nuovo Rete!).
Table 2 shows a clear editorial tendency to place the imperfect at level A2.

Syllabus	**Imperfect**	**CEFR level**
Chiaro!	Forms and use of the imperfect; use of the perfect and imperfect tenses.	A2
Domani	Imperfect indicative; perfect and imperfect.	A2
Nuovo Rete!	Imperfect indicative.	A2

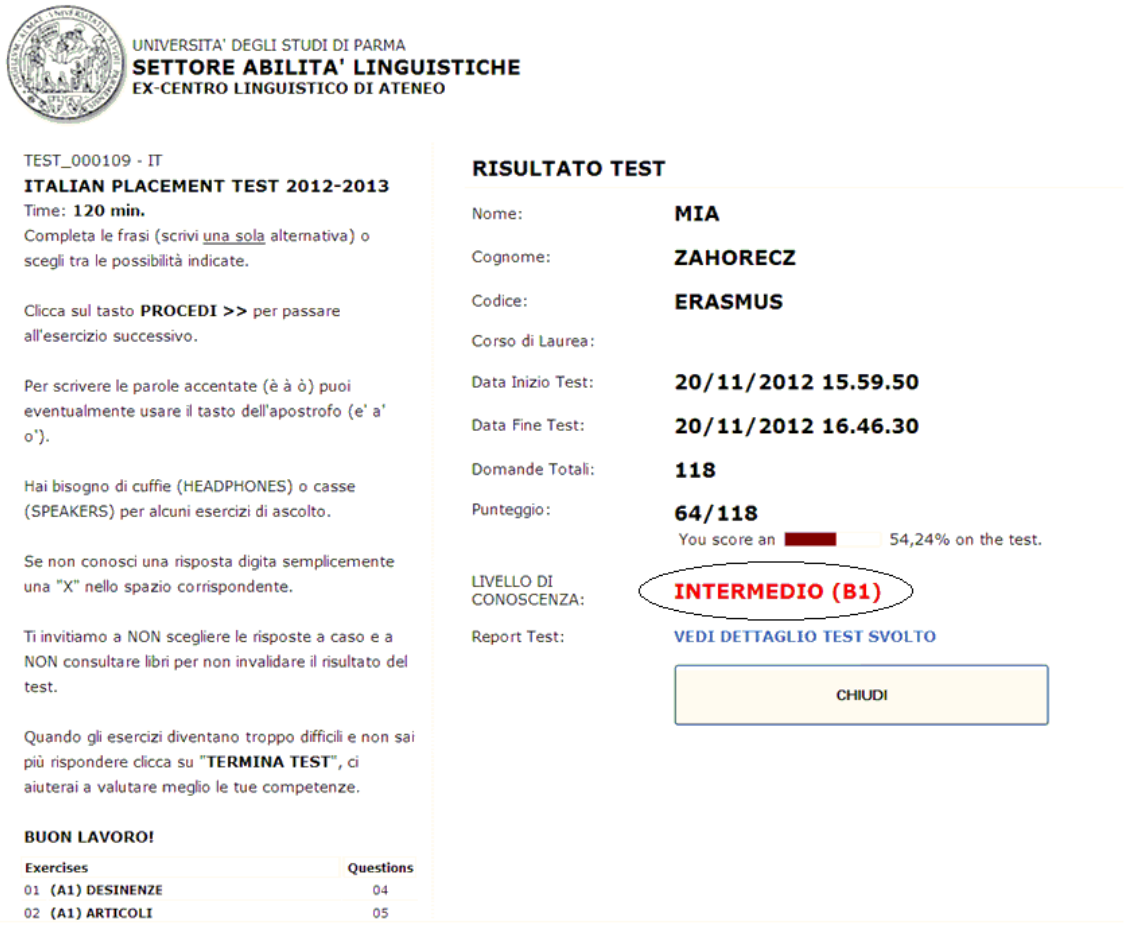
Table 2 – Indexing of perfect and imperfect tenses in manuals

As we couldn't ignore that trend, we decided to class the imperfect as A2.
We use this paradigmatic model every time we had a problem in tagging forms, structures or audio inputs.

The format of the new CEFR-aligned "Parma Placement" consists of six sections (A1-C2) divided into three subsections: the first one is dedicated to the use of the language, the second to the reading comprehension and the third to the listening comprehension.
After having piloted the new entry test with a restricted group of foreign students, we administered it to 242 Erasmus and defined the following cut-offs based on the percentage of correct answers:

Percentage of correct answers	CEFR level
from 0% to 36%	A1
from 37% to 48%	A2
from 49 to 71%	B1
from 72 to 88%	B2
from 89 to 96%	C1
from 97% to 100%	C2

As can be observed in figure 1, the system assigns students a CEFR level at the end of the test.


UNIVERSITA' DEGLI STUDI DI PARMA
SETTORE ABILITA' LINGUISTICHE
EX-CENTRO LINGUISTICO DI ATENEO

TEST_000109 - IT
ITALIAN PLACEMENT TEST 2012-2013
Time: 120 min.
Completa le frasi (scrivi una sola alternativa) o scegli tra le possibilità indicate.

Clicca sul tasto PROCEDI >> per passare all'esercizio successivo.

Per scrivere le parole accentate (è à ò) puoi eventualmente usare il tasto dell'apostrofo (e' a' o').

Hai bisogno di cuffie (HEADPHONES) o casse (SPEAKERS) per alcuni esercizi di ascolto.

Se non conosci una risposta digita semplicemente una "X" nello spazio corrispondente.

Ti invitiamo a NON scegliere le risposte a caso e a NON consultare libri per non invalidare il risultato del test.

Quando gli esercizi diventano troppo difficili e non sai più rispondere clicca su "TERMINA TEST", ci aiuterai a valutare meglio le tue competenze.

BUON LAVORO!

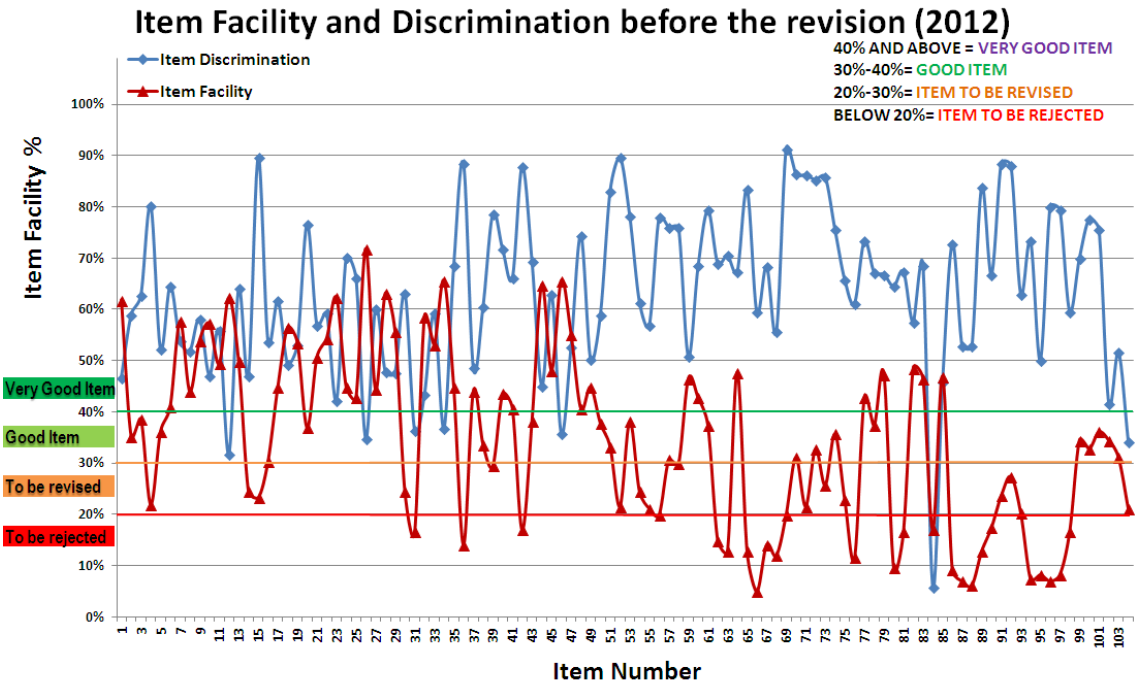
Exercises	Questions
01 (A1) DESINENZE	04
02 (A1) ARTICOLI	05

RISULTATO TEST

Nome: MIA
Cognome: ZAHORECZ
Codice: ERASMUS
Corso di Laurea:
Data Inizio Test: 20/11/2012 15.59.50
Data Fine Test: 20/11/2012 16.46.30
Domande Totali: 118
Punteggio: 64/118
You score an 54,24% on the test.
LIVELLO DI CONOSCENZA: INTERMEDIO (B1)
Report Test: VEDI DETTAGLIO TEST SVOLTO
CHIUDI

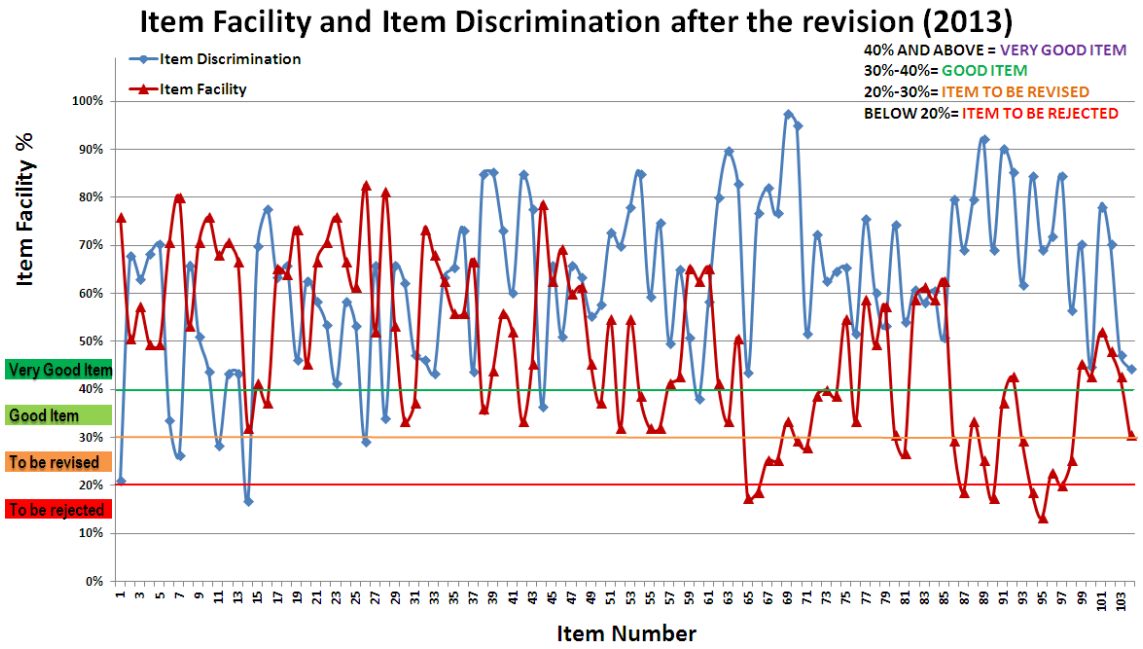
The survey that has been conducted among teachers, resulted in the test being appreciated: the levels assigned were realistic and useful to form more homogenous classes.
Since the "Parma Placement" worked well in the context it was used, we decided to prove its validity at an international level as well.
We then asked the University of Falun (Sweden) and the University of Zagreb (Croatia) to pilot it.
After the first administration, teachers from both Universities greatly appreciated the test results, as they proved to correspond to the real level of their students.
Since Falun and Zagreb asked to continue using it, we were encouraged to conduct an item analysis to improve the quality and accuracy of the items and to make sure that our placement was measuring at the CEFR levels it claimed it was.

Among the range of statistics on the performance of the items and the test as a whole, we took into account the "Item Facility" and the "Item Discrimination" indexes.
Generally speaking, for the IF and ID, values should be +.40 and above, although it is often necessary to accept lower values.
In our context, for example, we set out rules of thumb to detected "very good" and "good" items from items that needed "to be revised" or item that had to be "rejected".
A "very good" item scores 0.40 and above, a "good" item is ranked between 0.30 and 0.40, an item that needs "to be revised" between 0.20 and 0.30, and an item that had to be "rejected" below 0.20.
Although Graph 1 shows a general good IF and ID indexes, some areas below 0.20 needed improvement.



Graph 1 – Item Analysis before the revision (2012)

As far as the IF concerns, the items 67, 88, 94, 95 etc… proved to be too easy.
In order to improve them we either revised the stem or the difficulty level of the distracters.
As far as the ID concerns, we know that a good item discrimination should distinguish well between weaker and stronger test takers.
A highly discriminating item has an index approaching +1.00 (the closer 100% the better), showing that the strongest test takers are getting the item correct while the weakest are getting it wrong. On a good test, most items will be answered correctly by 0.30 to 0.80 of the examinees.
In our case, the item 85 proved not to discriminate between good and bad students because the best test takers got it wrong, while the worst got it right. Scoring below 0.05 we replaced it by a different one.
As can be observed in Graph 2 the item analysis we conducted after the revision of the placement on 54 new test takers, highlights a general improvement of the whole test.



Graph 2 – Item Analysis after the revision showing a general improvement .

Being aware that these results need to be further analyzed and interpreted, the "Parma Placement" proved to be:

- satisfactorily aligned with the CEFR levels;
- reliable and valid either for both Germanic/Slavic and Romance linguistic areas, in particular for Spanish-speaking students;
- practical and operational in terms of performance which has been improved by the item analysis that has been conducted.

We are committed to putting a great deal of effort into regularly checking that the items correspond to the CEFR levels on the basis of modern docimology and into the reconsideration of the cut-offs.
Since we are interested in further researching into placement tests, we take this opportunity to invite other Universities to administer the on-line "Parma Placement" asking us for a free delivering.
We would welcome - in exchange - a concrete feedback on the results being achieved.

References

Balboni, P. E. (1998). Tecniche didattiche. Italiano, lingue straniere e lingue classiche. Turin: Utet.
C-Test, http://www.c-test.de/deutsch/index.php?lang=de&section=ctest.

De Savorgnani, G. & Bergero, B. (2010). Chiaro! Corso di italiano. Livello A1. Florence: Alma.

Forapani, D. (2013). Test di ingresso e validità. Proposte per un miglioramento della performance del placement di italiano L2/LS per studenti ispanofoni, Language Learning in Higher Education, De Gruyter-Mouton, Berlin-New York, 115-127

Forapani, D. (2011). The tapestry of a placement test. Research into designing an Italian placement test for Erasmus students (who are native speakers of Romance languages), Tuttitalia, 40, 3-6

Freddi, G. (1994). Glottodidattica. Fondamenti, metodi e tecniche. Turin: Utet.

Hudson, T. & Clark, M. (Ed.). (2008). Case studies in foreign language placement: Practices and possibilities. University of Hawaii at Manoa: National Foreign Language Resource Center.

Lo Duca, M. G. (2006). Sillabo di Italiano L2 per studenti universitari in scambio. Rome: Carocci.

Mezzadri, M. & Balboni, P. E. (2010). Nuovo Rete! Livello A1. Perugia: Guerra.

Micheli, P. (Ed.).(1994). Test d'ingresso di italiano per stranieri. Rome: Bonacci.

Naddeo, C. M. & Guastalla, C. (2010). Domani. Corso di italiano. Livello A1. Florence: Alma.

Porcelli, G. (1994). Educazione linguistica e valutazione. Turin: Utet

Società Dante Alighieri, Syllabus Certificazione PLIDA: http://www.plida.it/plida/images/stories/documenti/sillabo.pdf (site visited 22.10.2012).

Università per Stranieri di Siena, Syllabus Certificazione CILS: http://www.gedi.it/cils/file/5/12/file/testo_linee_guida.pdf (site visited 22.10.2012).

Zdenka Gadušová & Andrea Billíková

Constantine the Philosopher University, Nitra, Slovakia

zgadusova@ukf.sk - abillikova@ukf.sk

English Tests for Secondary School Leavers in Slovakia

Bio data

Zdenka Gadušová is an experienced university administrator, project coordinator, researcher and dedicated teacher trainer in the field of methodology of foreign language teaching. For several years she worked in the position of the Head of the Department of English and American Studies where she co-ordinated several TEMPUS, Leonardo da Vinci and Comenius projects; at present she is the coordinator of Erasmus Mundus project. In 1996 she became the vice-dean for education and from 2002 to 2010 she was the Dean of the Faculty of Arts, UKF. Currently she is in the position of vice-dean for research. She is the author of three monographs and a number of articles, and a regular presenter of the results of her research work at conferences and seminars.

Andrea Billíková has been teaching at the Department of English and American Studies, Faculty of Arts, Constantine the Philosopher University in Nitra where she teaches courses on methodology and psycholinguistics. Her areas of interest are applying drama techniques as a principal learning, teaching and training tool for language learners and future English teachers. She closely cooperates with in-service English teachers via drama workshops and Drama Festival which is annually organised for primary, secondary and language schools in Nitra. She is also the head of the team of authors of the YES! series dedicated to secondary school leavers in Slovakia and the Czech Republic.

Abstract

Language policy in Slovakia follows the adopted European trends in language teaching and learning – any secondary school leaver should be proficient in two foreign languages apart from their mother tongue. Thus, all secondary school leavers in Slovakia are obliged to take, among other "Maturita exam" subjects, a final exam also in a foreign language either at B1 or B2 levels according to the CEFR. In our paper, we will present the concept of the final exam in English language taken as one of the "Maturita" subjects. We will describe its components and testing techniques. Specific attention will be paid to the process of students' preparation for the oral part of the exam, which secondary school leavers usually undergo either within English classes at school or on their own. The recent survey and research findings led us to create tailor-made study material for Slovak "Maturita" leavers that is based on reflective learning. During the presentation the material will not only be discussed (its aims and tasks) but also a video-recording showing the way students should be trained for the oral exam will be demonstrated.

CEFR and foreign language policy in Slovakia

The Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) was designed to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabi and curricula, the design of teaching and learning materials, and the assessment of foreign language proficiency. It describes foreign language proficiency at six levels and provides a basis for recognising language qualifications and thus facilitating educational and occupational mobility.

During the several decades of its use not only in Europe but currently we can say world-wide it has become one of the most widely used documents in the business of teaching and learning foreign languages. Slovakia is not an exception. In 2001 it joined the agreement of Eastern and Central European countries to have a common concept of final secondary school leaving examination in foreign languages at two levels of language proficiency under the joint name Independent Learner. This means that the learner should show conscious effort not only within English classes at school but also in out-of-school learning activities relevant to their interests and professional orientation. The learner should personally decide to what extent he or she will need and use English in the future and be responsible for the decision regarding the level of the examination he or she wants to sit.

The CEFR has been translated into Slovak and has widely been applied in the development of new curricula and syllabi for all levels of foreign language education in the Slovak system of education. In official documents it is stated that Slovak pupils have to achieve level A1 by the end of ISCED 1; level A2 by the end of ISCED 2; level B1 or B2 by the end of ISCED 3, depending on the type of the upper secondary school – vocational (B1) or grammar (B2). However, there is still quite a lot of discussion as to whether it is achievable within the allocated number of foreign language lessons in curricula and with all learners. The situation in Slovakia is very similar to the survey carried out in spring 2011 in 14 European countries, which revealed that Europeans still need to improve their knowledge of foreign languages, as there is a wide range of ability across the participating countries.

Having established the goals to be achieved in foreign language education, the tools for testing declared achievements have been developed as well. National testing of foreign language competence of secondary school graduates has more than a decade-long history in Slovakia. At first, the testing tools, developed for three levels of proficiency (A, B, and C – among them A as the highest and C as the lowest), were piloted in a number of schools for several years (2000 – 2006); then, in 2007 the first national piloting of foreign language testing at two levels (A and B) was carried out and since 2008 organisation of the secondary school leaving examination has been codified by legislation – Act N° 245/2008 on education (the so-called School Act) and Decree 318/2008 on completion of secondary school education. In 2008 the tests still were marked as A (higher level) and B (lower level) and only since 2009 have they been marked as B1 and B2 according to the language levels specified by the CEFR. Learners were free to choose either of them for four years but since the school year 2011/2012 a new amendment to the previous Act was introduced which says that all grammar school students must pass the examination at the level of B2 and students from other secondary schools can still decide which level of the exam they would like to pass.

Concept of secondary school final exam in English

The idea behind the new model of the final secondary school exam in English was to have common requirements for language knowledge and skills of learners, to have national evaluation criteria for assessment of language competence of learners and to have an as objective as possible final exam for secondary school leavers. The new model of this exam should enable secondary school graduates to study not only in Slovakia but also abroad and have better employment opportunities in the European labour market.

The secondary school leaving examination, the so-called "Maturita", in foreign languages is based on the national document Secondary School Graduate Target Requirements, which were developed and closely match CEFR standards. The exam consists of two parts – external and internal, the internal having a written part and an oral one. As the external part of the exam and the writing test of the internal part are carried out nationally, the institution appointed responsible for both of them is the National Institute for Certified Measurements in Bratislava.

The Institute is responsible for the development of standardised tests for all six foreign languages (English, German, French, Russian, Spanish and Italian) at B1 and B2 levels and it is also responsible for tests distribution, statistical assessment and analysis, and evaluation.
The oral form of the internal part of the "Maturita" exam is organised by each upper secondary school in close cooperation with regional school offices.

The external part of the "Maturita" exam is a written test. Its content is based on Secondary School Graduate Target Requirements on learners´ foreign language knowledge and skills either at the B1 or B2 level of the CEFR. The external part of the "Maturita" exam, whether for the B1 or B2 level, has three parts:

- listening comprehension
- grammar and vocabulary
- reading comprehension

As stated above the test verifies graduates´ foreign language competences in listening comprehension, reading comprehension and language in use. Topics are specified in Target Requirements and vary in different parts of the test.

Possible testing techniques used in the tests can be:

- short answer items
- multiple choice items (either with 3 or 4 distracters)
- matching tasks
- gap filling or substitution tasks
- ordering tasks
- true / false tasks
- filling in the correct form of a verb or word
- cloze test.

The Listening Comprehension paper in the external part of the "Maturita" exam usually includes three different (as to their topic, length and form – dialogue, monologue, speakers – female, male) recorded texts which are played twice and comprehension of each of them is tested by three different testing techniques, each of them with 6 or 7 items to be solved.

The Language in Use paper is the only one which differs in the number of items and timing for different levels of proficiency – for B1 it has 20 items (to be completed in 25minutes) and for B2 – 40 items (in 45minutes). The testing techniques are also here text-based and the paper has either 2 (B1) or 3 (B2) parts.

The last part of the test – the Reading Comprehension paper, consists of three topic-based texts for checking comprehension, in which different testing techniques are used. To complete this section, 45 minutes are allocated.

Since the school year 2007/2008 the external part of the "Maturita" exam (written test) has been carried out regularly in March so that its results would be available during the oral part of the "Maturita" exam in May. Currently, we can compare the results achieved in these tests for the last five years (2008 – 2012); for this year (2013) they are not available at the moment. The collected data show the situation at a national level, as all secondary schools in Slovakia in which the students take the "Maturita" exam have been included in testing. This means that each year about 40 000 secondary school leavers were tested and the total number of tested students for the five-year period is 198 008 students, out of which 163 027 students passed the B1 tests and 34 981 students (17,6% out of the total number) passed B2 tests. (see the table below)

The tests were designed to match the standards in the CEFR for B1 and B2 levels. But as the table shows the overall results achieved in testing are not very satisfying. The results in B1 tests show just a semi-success, as the overall results are only slightly above 50% and in the last year (2012) even below 50%. This can be partially connected with the fact that all grammar school leavers had to pass B2 tests in 2012 (which means that some better students, who would have chosen the B1 test, had to pass the B2 test) and this also had a direct consequence on the results achieved in the B2 tests in 2012 which were much worse (55,4%) than in the previous years when they ranged between 65,1% and 70,5%.

Year	Level	Number of students	SUCCESS			
			Listening comprehension	Language in use	Reading comprehension	Total
2008	**B**	26 255	62,2 %	63,2 %	60,3 %	61,9 %
	A	6 753	78,4%	54,6%	78,7%	70,5%
2009	**B1**	34 312	65,6 %	40,1 %	44,7 %	50,1 %
	B2	5 356	68,8 %	58,2 %	61,7 %	65,1 %
2010	**B1**	37 939	54,9 %	55,5 %	52,1 %	54,2 %
	B2	3 953	74,3 %	60,1%	76,4 %	70,3 %
2011	**B1**	38 198	66,2 %	48,6 %	53,9 %	56,2 %
	B2	3 268	76,7 %	62,1 %	70,1 %	69,6 %
2012	**B1**	26 323	58,7 %	39,4 %	47,0 %	48,4 %
	B2	15 651	55,8 %	52,7 %	57,6 %	55,4%
2008 - 2012	**B1**	**163 027**	**61,5 %**	**49,4 %**	**51,6 %**	**54,2 %**
	B2	**34 981**	**70,8%**	**57,5%**	**68,9%**	**66,2%**

Table 1: Overview of testing results for the external part of "Maturita" exam in Slovakia (2008 – 2012)

As regards the different sections of the "Maturita" tests, the results show that secondary school leavers have the biggest problem with the completion of the Language in Use paper at both levels (B1 and B2), i.e. with the accuracy of language use in comparison with comprehension of the target (English) language texts. In B2 tests we can see very tiny difference between reading comprehension and listening comprehension achievements. The students are equally good at both reading and listening. On the other hand the results of B1 tests show surprisingly better achievements in listening comprehension (a total of 61,5% for the 5 years) than in reading comprehension (a total of 51,6% for the 5 years). The reasons for this we see in two trends. One of them is low attention paid by students to accuracy of expression and reading – students are not focused enough on what they read and, in general, are not very keen on reading nowadays. The other current trend is the use of more up-to-date course-books with many audio recordings and supplementary materials as well as with many possibilities for students to listen to audio texts and performances out of school, in real life. All this undoubtedly has a positive impact on the development of students´ listening skills.

Another part of the "Maturita" exam is the so-called written form of the internal part of the exam in which the writing skills of students are tested. To complete this test students have 60 minutes. It is called the "internal part" as "internal" English teachers from the schools assess these tests according to the national set of four criteria: content of the text, structure and paragraphing of the text, grammar and vocabulary. For each of these criteria 5 points are allocated which makes a total of 20 points.

Students taking this test at both levels are always given just one topic to write about in a set genre (B1: 160 – 180 words and B2: 200 – 220 words) which is not always easy for them to do, as they are often not trained in how to approach the process of writing and how to develop

the topic. Very often they are just told the topic and are asked to deliver its written form in a certain genre. Furthermore, not having a choice of topic or genre makes it even more difficult for students. To demonstrate some of the topics and genres we can state some recent examples of such tasks for:

B1:
2012 – topic: My School - Write a letter to your English penfriend about the school you attend.
Include: two things you like about attending it; two things you would like to change and how.
2013 – topic: My Favourite Book - Write a description of a book you like very much.
Include: the title and the writer; the reasons for your choice; a short description of the plot; the characters you admire in the book; why you would recommend it to your best friend.

B2:
2010 – topic: A Birthday Party - Write an article for an English language school magazine about a birthday party you attended last weekend.
Within the topic consider: the reasons for attending the party; the location and decorations, the meals and drinks provided; the people you met there; the atmosphere at the party.
2012 – topic: An Ideal School - Write an essay about an ideal secondary school for you.
Justify your opinions about: the setting you would study in; relationships between teachers and students; the subjects you would study; testing; extra-curricular activities.

The achieved results in this part of "Maturita" tests are, however, not significantly worse or better. They match quite well the results of students achieved in the external part of the "Maturita" exam, though, we can say that in each of the five years, they were slightly higher, as e.g. last year (2012) at B1 level the total for the whole set of external tests was 48,4% and for writing tests it was 54,6% and at B2 it was 55,4% to 71,8 %. The better results of the writing tests can be attributed to the fact that they are evaluated internally. As for the four criteria which are applied to assess these tests, we can say that the results in the first two criteria (content of the text, structure and paragraphing of the text) are always higher than in the other two criteria (grammar and vocabulary), which again is probably connected with the lack of attention paid by students to accuracy of expression.

Oral part of the 'Maturita' exam – process-oriented approach

Before taking the final exam in English language called "Maturita", most secondary school leavers undergo exam preparation either in their English classes with their teachers or at home. Besides three to four regular English classes allotted per week[1], "Maturita" students are recommended to take extra classes of "English conversation" where the main emphasis is put on the preparation for the oral "Maturita" exam (OME)[2]. These lessons are optional therefore not necessarily all learners go through systematic guidance

[111] The number of English lessons depends on the type of secondary school. Secondary grammar schools which prepare their students for B2 proficiency level offer 4 lessons a week, while so-called vocational schools preparing either for B2 or B1 make do with 3 lessons per week of regular classes and 2 lessons of optional conversation classes. Finally, for so-called "joined secondary schools" (preparing students for professions such as hairdresser, car mechanic, cook, etc.) there are 2-4 lessons allotted per week with 1 conversation lesson taken on an optional basis.

[2] OME- the abbreviation standing for oral "maturita" exam. It is the external part of the final exam taken at the end of the last year at secondary schools.

provided by their teachers. Preparation for this exam is based on mastering 25 topics, which are the same for all foreign languages and for both proficiency levels. The difference is in the amount and difficulty of the required and acquired vocabulary. The national document mentioned earlier[3] describes the required abilities, skills and competencies needed for completing the following communication tasks that are tested during OME: 1. description of visual stimuli (picture, photography, graph), 2. expression of personal opinion, 3. simulation or role-play. The main difference between B1 and B2 lies in the choice of vocabulary, grammar structures, content of language performance and ability to react promptly. The evaluation of the oral performance takes into account the choice of grammar structures, vocabulary, knowledge of facts, information and provision of arguments and opinions. Whether "Maturita" students are really heading towards meeting the above mentioned requirements [4] and whether they are getting ready for the OME or not is, besides many other important factors, we believe, conditioned by the type of applied approach during the exam preparation process which "Maturita" students undergo either at schools or at home.

Preparing students for the OME is undoubtedly a demanding and lengthy process. The quality of students' performance during OME is not only conditioned by language competence and time spent on preparation for the exam but also by presence of personal, affective and socio-cultural factors such as fear, anxiety, concentration, mood of speakers and listeners, their interest and need to communicate, the presence of other participants, their mutual relationship, social status, the importance of the communication situation, topic, communication task and many others. Last but not least, what is important is the awareness and application of appropriate learning and communication strategies used, depending on the communicative situations and tasks. Referring to the CEFR, "communication and learning involve the performance of tasks which are not solely language tasks even though they involve language activities and make demands upon the individual's communicative competence. To the extent that these tasks are neither routine nor automatic, they require the use of strategies in communicating and learning." (CEFR, p. 15) This supports our belief that besides regular practice of the target language and teachers' professional instruction on successful accomplishment of OME, both learning and communication strategies are equally important in the implementation of the OME preparation process.

Strategies in general are seen as helpful tools for language user to "mobilize and balance his or her resources, to activate skills and procedures, in order to fulfil the demands of communication in context and successfully complete the task in question in the most comprehensive or most economical way feasible depending on his or her precise purpose." (ibid p. 57) Communication strategies are defined as useful verbal and nonverbal tools that are used to cover communication difficulties that happen in everyday communication due to insufficient language knowledge or presence of performance variables. They are seen from two different angles: either as "tricks" to hide poor language competence used by less competent language users or "useful tools" to make communication smooth and efficient.[5] Heading towards meeting the requirements discussed above, we can identify with the opinion stated in the CEFR that "Progress in language learning is most clearly evidenced in the learner's ability to engage in observable language activities and to operate communication strategies. (ibid, p. 57) In order to find out the relevance of importance of learning and communication strategies in the preparation process for OME, the survey on learning strategies and research on communication strategies were carried out among ex-Maturita takers at the Department

[3] Secondary School Graduate Target Requirements is the Slovak national document aiming to provide the requirements for specific language proficiency levels A1, A2, B1, B2 matching CEFR standards.
[4] more on requirements for B1 and B2 proficiency levels see the national document at http://www.nucem.sk/sk/maturita (the text is available in Slovak)
[5] The definitions, study and research on communication strategies come from the unpublished doctoral dissertation by Andrea Billíková, Comenius University, 2007.

of English and American Studies, Constantine the Philosophy University in Nitra, Slovakia.

The recently conducted survey (2013) among fifty respondents indicates that besides English textbooks, various supplementary materials (such as magazines, newspapers, simple reader books) and on-line study sources (English learning websites and on-line videos) were used either at school or at home when getting ready for OME. As for learning strategies, the findings were quite striking: memorizing isolated topic-based vocabulary and repeating texts from course-books were predominantly used. These findings reveal persisting traditional learning strategies applied in oral "Maturita" preparation which dominated some decades ago in Slovakia when "Maturita" leavers normally memorised texts and vocabulary related to "Maturita" topics on social and cultural life, and historical and geographical facts about English-speaking countries. This was the most shocking finding for us since we expected that a more process-oriented approach[6] would be currently used by English teachers and learners who are getting ready for their OME. Apart from repeating "Maturita" texts in front of the class, just a few respondents said they were guided on how to analyse and assess the quality of their peers' oral performance while preparing to complete the oral "Maturita" tasks.[7] Despite persisting attempts to foster learners' autonomy via a process-oriented approach, very little attention was paid to the self-assessment and self-reflection that we consider to be crucial elements in the process of preparation for the OME. The survey results show that in spite of teachers' guidance and instruction, the trend to apply the traditional, product-oriented approach[8] dominates over the process-oriented approach in oral "Maturita" preparation.

The earlier quantitative-qualitative research[9] carried out in the years 2000-2007 among 38 ex-Maturita leavers aimed at examining the relevance and teachability of communication strategies in a formal setting by using drama techniques in order to foster communicative competence (specifically strategic sub-competence) and performance. We tried to find out whether students after direct training in communication strategies would be better able to cope with their language difficulties by using a variety of compensation strategies in an efficient and comprehensible way in different communication situations and whether they would become more self-confident language users. As for the research methods, the students´ questionnaires, direct observations, experimental teaching and students' journals were used. Students in the experimental group underwent a complete communication strategies training including compensation, interaction-modification and conversation strategies via drama techniques. It has been confirmed that teaching communication strategies in a foreign language is possible to some extent. We found out that the choice of compensation strategies depends also on communication tasks and forms of communication. Drama techniques were appropriate teaching and learning tools to increase strategic awareness and overcome psychological barriers our students used to face in the process of communication.

The findings from the previously mentioned survey and research motivated us to design process-oriented study material for OME preparation[10]. Ready-made video recordings of all (simulated) varieties of OME tasks completed with success or difficulties lead their viewers through the process of detailed analysis and evaluation of the oral performance of

[6] see also the Chapter on the Action-oriented approach in CEFR, p. 15

[7] Oral "Maturita" tasks are the same for B1 and B2 maturita takers: 1. describe the visual stimuli 2. express your opinion, 3. do simulation or role-play.

[8] Traditional approach for oral "maturita" preparation is in our perception equal to the product-oriented approach which ignores learner independence, hence memorising vocabulary and texts bounded to maturita themes are the main (and often only) learning strategies.

[9] The research was part of the doctoral dissertation of A. Billíková that was defended in 2007 at Comenius University in Bratislava, Slovakia.

[10] The mentioned study material is part of the book by Billíková, A. , Kondelová S. (2012). *YES! Angličtina-maturita-vyššia úroveň (B2).* Nitra: Enigma. It is also recommended by the Ministry of Education in Slovakia.

"Maturita" takers. The accompanying activities encourage learners to analyse the quality of Maturita task completion, recognise communication breakdowns of observed oral Maturita takers and suggest possible ways for their remediation. Furthermore, video viewers can compare and contrast their evaluation with recorded teacher's comments on the efficacy of applied learning and communication strategies in task fulfilment. After going through detailed analysis, evaluation and reflection on fulfilling all OME tasks recorded on the video, "Maturita" takers should be able to better succeed at OME with appropriate application of both learning and communication strategies irrespective of the preparation for OME that happens at school or at home.

Conclusions

Preparing secondary school leavers for their "Maturita" exam in English and relevant testing of the level of their language proficiency is a very demanding and responsible task for teachers. As the results in both oral and written "Maturita" tests in Slovakia have shown for the last five years, it is very important to apply the process-oriented approach in getting the learners ready to meet the stated requirements adopted from the CEFR for the "Maturita" exam. Though the numbers pointing to the success of secondary school leavers with a "Maturita" exam may not seem to be very impressive, if partial achievements are taken into consideration, we can see that secondary school leavers show quite satisfactory performance in comprehension of language despite some drawbacks in language use. This leads us to the conclusion that the standards stated in the CEFR should probably be revised from the point of view of the authenticity of language use, whether it is used for comprehension or expression of one's own ideas.

References

Cieľové požiadavky na vedomosti a zručnosti maturantov z anglického jazyka úroveň B1. (2008). Bratislava: Štátny pedagogický ústav. Retrieved from http://www.nucem.sk/sk/maturita.

Cieľové požiadavky na vedomosti a zručnosti maturantov z anglického jazyka úroveň B2. (2008). Bratislava: Štátny pedagogický ústav. Retrieved from http://www.nucem.sk/sk/maturita.

Maturita 2008, 2009, 2010, 2011, 2012, 2013. Retrieved from http://www.nucem.sk/sk/maturita.

Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe. Retrieved from http://www.coe.int.

Billíková, A. (2007). Rozvoj a nácvik komunikačných stratégií v ústnom cudzojazyčnom prejave pomocou dramatických techník. (Unpublished doctoral dissertation). Comenius University, Bratislava.

Billíková, A. & Kondelová S. (2012). YES! Angličtina – maturita - vyššia úroveň (B2). Nitra: Enigma.

< The School Act > <2008> < Act N° 245/2008> (SK).

Jesús García Laborda, Mary Frances Litzler & Marian Amengual Pizarro

Universidad de Alcala, Alcala-Madrid, Spain

jesus.garcialaborda@uah.es

Can Spanish High School Students Speak English?

Bio data

Jesús García Laborda is an associate professor at Universidad de Alcala (Madrid, Spain). Dr Garcia Laborda has a PhD in English Philology and an EdD in Language Education. His current research covers many areas of compluter implementations for language learning and testing along with ESP and teacher training especially for the Spanish University Entrance Examination or the implications of implementing such test in teacher training along with more traditional approaches to teacher education and their development of both cognitive and computer skills. His publications include papers in Computers & Education, British Journal of Educational Technology or Educational Technology and Society.

Mary Frances Litzler, PhD, teaches English at the University of Alcala and the British Council in central Spain. At the university level she teaches English language and linguistics in the degree programs for English Studies and Modern Languages and Translation. Her research interests are new technology in foreign language learning, language testing, and medieval text editing. She has published in Procedia - Social and Behavioral Sciences, International Educational Studies, Language Learning and Technology, and Porta Linguarum, among others. Her PhD thesis was an edition of a corpus of 15th century medical prologues in English. She has taught English for Academic Purposes in the United States, Spain and Japan.

Marian Amengual-Pizarro currently holds the post of lecturer in the Department of Spanish, Modern and Classical Philology (English Philology) at the University of the Balearic Islands. Her research has been mainly in the area of testing. She has been coordinating the English Test in the Spanish University Admission Examination at the University of the Balearic Islands (UIB) since 2003. Her other main research area is the teaching of English as a foreign language. She has been the 'Language Teaching' Panel coordinator at AESLA (Spanish Association of Applied Linguistics) for eight years (2004-2008). She is also the current secretary of AEDEAN Association (Spanish Association of Anglo-American Studies), and member of the editorial board of the journal "e-resla" and "Revista de Educación". She has published in national and international refereed journals on applied linguistics and education.

Abstract

Spain will go through a number of educational changes in the next six months because a new educational reform will be implemented from the academic year 2013-2014. Similarly to the "No Child Left Behind Act", it is thought that assessment will determine the specific support to certain educational programs and school. Assessment is considered to have a great impact in the initiatives for improvement in the overall underachieving educational system (PISA 2012, OCDE indicators 2012, European Survey of Language Competence). When planning the most significant proposals to implement the educational system, the researchers of the OPENPAU project considered the importance of implementing large scale assessments for languages with limited

resources. Since computer labs in high schools are neither secure nor updated and do not usually have enough posts, proposals for alternative solutions have been suggested. In that sense, ubiquitous alternatives provide options to increase the number of test candidates taking the test simultaneously. We also considered the demands of teachers and educational boards in test design. Accordingly, we considered that test takers should be able to be tested in speaking, listening, reading and writing. As a consequence, the informatics team considered two main proposals: 1) The design of a specific test to implement speaking exams as the OPI or the speaking sections of TOEFL; 2) to use ubiquitous m-design that can be used in a number of devices such as i-pads, mobiles or tablet PC. The test validity followed the validation standards by Weir (2005). The presentation describes the technological and practical aspects of delivery and application. The paper also proposes a research agenda including different applications from the Baccalaureate General Test.

Short paper

The language problems evidenced by the European Survey on Language Competences (ESLC)[1] have a direct impact in the concerns of the Spanish Ministry of Education, Culture & Sports (MECD). The fact that Spanish students are among the worst English learners needs aserious revision in the prospective educational reform. PISA and other international assessments are also going online and Spain has proved that Spanish students in high school are far behind most European countries. If fact, table 1 shows this diadvantaged position agains most of the other European countries.

Educational system	Language	Reading			Listening			Writing		
		Pre -A1	A	B	Pre -A1	A	B	Pre -A1	A	B
Bulgaria	English	23	43	34	23	37	40	15	52	32
Croatia	English	16	44	40	12	32	56	5	49	45
Estonia	English	7	33	60	10	27	63	3	37	60
Flemish Community of Belgium	French	12	63	24	17	62	20	19	59	22
France	English	28	59	13	41	46	14	24	61	16
French Community of Belgium	English	10	59	31	18	55	27	6	65	29
German Community of Belgium	French	10	52	38	11	49	40	8	51	41
Greece	English	15	40	45	19	35	46	7	41	53
Malta	English	4	17	79	3	11	86	0	17	83
Netherlands	English	4	36	60	3	21	77	0	39	60
Poland	English	27	49	24	27	45	28	19	59	23
Portugal	English	20	53	26	23	39	38	18	55	27
Slovenia	English	12	42	47	5	28	67	1	51	48
Spain	English	18	53	29	32	44	24	15	58	27
Sweden	English	1	18	81	1	9	91	0	24	75
UK England	French	22	68	10	30	62	8	36	54	10

Table 1. First foreign language - percentage of pupils achieving broad levels by skill and educational system (source: European Survey on Language Competences)

Given this situation, the Spanish Ministry of Education, Culture & Sports (MECD) is aware of the following factors:

1. There is a need to revise the educational paradigm in Spain including language policies;

[1] http://ec.europa.eu/languages/eslc/docs/en/executive-summary-eslc_en.pdf and http://ec.europa.ec/languages/eslc/docs/en

2. Assessment, evaluation and testing should increase its relevance in the new educational reform;
3. Testing may have both positive and negative influence so it needs to ne adequately shaped;
4. Periodical assessments may provide important benefits for schools and learining;
5. The impact of these assessments should be moderated and account for the socio-economic inequeslities in the Spanish society;
6. Not all the resources provided to schools should depend on the results of those assessments to avoid undesirable counter-effects as seen in policies such as the "No Child Left Behind" in the US[2].

Current research

In order to propose different alternatives for the current situation the OPENPAU project followed two main lines to address the analysis of current limitations of Spanish students. On the one hand, the coordinator of the project established lines of cooperation between the research project and the MECD. The general ideas was that the experience of tf the OPENPAU project served to provides ideas to improve the current situation and also to revise an internal report on the high school leavers' English proficiency. As a counterpart, the MECD offered to provide information on the current research through the online delivery of the research database. The starting point for this paper is precisely are these results. Simultaneously the OPENPAU project studied the speaking production of 150 first semester university students of different fields.

The report from the MECD

The report done by the MECD was done by teachers who do not rate the current University Entrance Examination but regular practitioners with limited experience in language testing and even more limited theoretical and practical knowledge in six Spanish regions. 1033 students were tested in a two part exam whose tasks were informal examiner-student conversation and the description of a picture. Results released by the MEC evidenced that more than half (61.3%) of the students were over the pass grade at a B1 level in the CEFR. In fact, the report did not find significant differences in reference to sex or age (Table 2).

	Criterio					Total
	Alcance	Corrección Gramatical	Fluidez	Interacción	Coherencia	
Parte 1ª	65.08	60.39	64.66	68.17	68.51	63.65
Parte 2ª	61.06	55.53	62.81	64.49	65.66	60.80
Total	60.22	54.77	61.89	63.82	64.57	61.39
pvalor 1ª-2ª	0.04183	0.01614	0.34711	0.05709	0.13836	0.15091

Table 2. Percentage of students who pass at the B1 level in the CEFR

[2] http://www.gallup.com/poll/156800/no-child-left-behind-rated-negatively-positively.aspx or Educational Research Newsletter. "Pros and Cons of NCLB: What the Research Says."*Educational Research | Education Training | ERNweb*. 2006. Retrieved 3/26/2013 from. http://www.ernweb.com/public/892.cfm

The report from the OPENPAU project

The study within the OPENPAU project was done with just 230 students (only the results of 106 are presented here) from five different regions. Raters were specialists who have taken part in the University Entrance Examination for, at least, two years. Instead of looking at the interviews from a criterion-referenced perspective, they approached the interviews with a norm-referenced one. They intended to place the students in their level in the CEFR. The research team observed the following (table 3).

	A1	*A2*	*B1*	*B2*	*C1*	*Lost*
Occurrences	26	33	29	10	3	2
Percentage	24%	31%	27%	9%	3%	

Analysis and discussion

Although the Spanish legislation requieres high school leavers to achieve a B1 competence level, we observed that there is a large amount of students who do not get it in the MECD study and, what is even worse, the study done within the OPENPAU project shows that Spanish students tend to underachieve. It is also significant the limited number of student who are at B2 or higher. There are a number of issues that are still under debate according to the results differences. First, was the study by the MECD reliable. We may consider that if textbooks and work in class is usually done in pairs, the test was not fair to the students. Then, why the high results? Second, may it be that the teacher's expectations about the students' performance actually led to have an influence on the grades? Could the issue of the rater's expectations have also had an impact on the more experienced raters. Since the speaking tasks have never been included in the University Entrance Examination, could they be considered as "experienced"?

Conclusions: Competence and performance

The initial question still remains open. There are three main aspects that shape our response:

- Practicality: Is it really practical a high school leaving exam? Is the CEFR the best way to assess the students' achievement in high school? What should be an acceptable proficiency level? The practicality has been jeopardized in these two studies. Most likely, students who do not need a specific grade to enroll in certain universities may not need to take a language test to see whether they can or cannot perform at the adequate level in college. The CEFR seems at this point a week tool to assess weaker students who may not have interest in international professional mobility. Like in other studies, this comparison may evidence significant differences in the study plan but it may also be that the benchmarking provided by the CEFR is not as clear as it has been suggested a number of times.
- Test purpose: The studies hereby presented show that a test's construct definition is one of the difficulties associated with the CEFR. The CEFR should be considered globally. Otherwise, we may be introducing overgeneralizations that mislead the actual language profile associated to a student. The limited set of tasks in the MECD study may not lead to such a negative situation as the one suggested by the use of the CEFR in the OPENPAU research.

Our conclusion in these first data obtained in the OPENPAU project is that the CEFR can be a valuable tool when the ideal conditions of global assessment are present. It can also be valuable to analyze in terms of large number of students but not so good when considering individuals. In our study and given the expected Spanish educational reform, the CEFR is a powerful ally to determine the realities and needs of the country and to implement new procedures of assessment such as dynamic evaluation (Pohner, 2008; Pohner and Lantolf, 2006). To conclude, given the results of these experiences, the Spanish students will probably be able to travel abroad and communicate their basic needs but they certainly are not ready to pursue higher education if communication in English is necessary. The key issue may be in a revision of how we assess in Spain or whether introducing high stakes testing will be beneficial for the educational system as a whole.

The researchers would like to express their gratitude to the Ministry of Research and Innovation of Spain (MICINN) for supporting the development and implementation OPENPAU research project (FFI2011-22442) with cofounding with ERDF funds under the 2008-2011 plan. Finally, this paper could not have been possible without the contacts with Matt Poehner and James Lantolf from Penn State University through the Research Mobility for Senior Reasearchers grant (with funding from the Spanish Ministry of Education, Culture & Sports, PRX12/00376).

References

Lantolf, J. P. & Poehner, M. E. (2006). Dynamic Assessment in the Foreign Language Classroom. A Teachers Guide. Center for Advanced Language Proficiency Education and Research, The Pennsylvania State University, University Park, PA.

Poehner, M. E. (2008). Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development. Berlin: Springer Publishing.

Weir, C. (2005) Language testing and validation: An evidence-based approach, Palgrave MacMillan.

Jesús García Laborda, Mary Frances Litzler, Teresa Magal Royo & Nuria Otero de Juan

Universidad de Alcala, Alcala-Madrid, Spain

jesus.garcialaborda@uah.es

Proposals of Ubiquitous Delivery of the Foreign Language Paper of the Spanish Baccalaureate General Test

Bio data

Jesús García Laborda is an associate professor at Universidad de Alcala (Madrid, Spain). Dr Garcia Laborda has a PhD in English Philology and an EdD in Language Education. His current research covers many areas of compluter implementations for language learning and testing along with ESP and teacher training especially for the Spanish University Entrance Examination or the implications of implementing such test in teacher training along with more traditional approaches to teacher education and their development of both cognitive and computer skills. His publications include papers in Computers & Education, British Journal of Educational Technology or Educational Technology and Society.

Mary Frances Litzler, PhD, teaches English at the University of Alcala and the British Council in central Spain. At the university level she teaches English language and linguistics in the degree programs for English Studies and Modern Languages and Translation. Her research interests are new technology in foreign language learning, language testing, and medieval text editing. She has published in Procedia - Social and Behavioral Sciences, International Educational Studies, Language Learning and Technology, and Porta Linguarum, among others. Her PhD thesis was an edition of a corpus of 15th century medical prologues in English. She has taught English for Academic Purposes in the United States, Spain and Japan.

Teresa Magal is a Full professor at Universidad Politécnica de Valencia where she teaches graphic design for educational purposes and the media. She has extensively researched in interface graphic design and educational computer testing architecture. She has published in Computers & Education, Eurasian Journal of Educational Research and Iberica The Journal of English for Specific Purposes.

Nuria Otero de Juan is a research assistant at Universidad de Alcala. She is currently beginning her career in international conferences in Europe.

Abstract

Spain will go through a number of educational changes in the next six months because a new educational reform will be implemented from the academic year 2013-2014. Similarly to the "No Child Left Behind Act", it is thought that assessment will determine the specific support to certain educational programs and school. Assessment is considered to have a great impact in the initiatives for improvement in the overall underachieving educational system (PISA 2012, OCDE indicators 2012, European Survey of Language Competence). When planning the most significant proposals to implement the educational system, the researchers of the OPENPAU project considered the importance of implementing large scale assessments for languages with limited resources. Since computer labs in high schools are neither secure nor updated and do not usually have enough posts, proposals for alternative solutions have been suggested. In

that sense, ubiquitous alternatives provide options to increase the number of test candidates taking the test simultaneously. We also considered the demands of teachers and educational boards in test design. Accordingly, we considered that test takers should be able to be tested in speaking, listening, reading and writing. As a consequence, the informatics team considered two main proposals: 1) The design of a specific test to implement speaking exams as the OPI or the speaking sections of TOEFL; 2) to use ubiquitous m-design that can be used in a number of devices such as i-pads, mobiles or tablet PC. The test validity followed the validation standards by Weir (2005). The presentation describes the technological and practical aspects of delivery and application. The paper also proposes a research agenda including different applications from the Baccalaureate General Test.

Short paper

Introduction

Language problems evidenced by the European Survey on Language Competences (ESLC)[1] have a direct impact in the concerns of the Spanish Ministry of Education, Culture & Sports (MECD). The fact that Spanish students are among the worst English learners needs a serious revision in the prospective educational reform. Direct measures, among others (such as increase in the bilingual education programs or a 25% increment of the number of teaching hours) include a revision of the diagnosis test policy currently followed. It has been suggested that adequacy of testing may benefit to a great extent the language teaching approach and methodology currently used in most of the classes. There is significant evidence that shows that positive washback in speaking (Alderson & Wall, 1993; Cheng, 2003; Hirai & Koizumi, 2009) especially for Spanish speakers (Munoz & Alvarez, 2010). However, although the cases of positive washback are numerous, specific preparation for a test may also lead to devote excessive time for preparation. In the Spanish case, Garcia Laborda and Fernandez Alvarez (2012) observed that while teachers would begin to prepare the writing tasks for a University Entrance Examination just two years before the test were held, if they had to do so for the speaking tasks they would begin up to four years before. From the data obtained through focus groups and a questionnaire they concluded that Spanish teachers are afraid for a number of reasons such as limited practice in the classroom, socio-economic factors, and others. In a different paper, Garcia Laborda (2012) suggested that if students had the adequate training through computer based speaking tasks Spanish students would improve their performance almost immediately. This position has also been supported by studies on synchronous communication (Vetter & Chanier, 2006; AbuSeileek,, 2007; AbuSeileek, 2012; Jenks, 2012), asynchronous communication (Hew & Cheung, 2012), programmed speech for presentations (Kunioshi, 2012). Researchers have also mentioned that coherence in education demands that is students use computers increasingly, they should also be tested through such a delivery means (Satar & Ozdener, 2008; Alderson, 2009; Wang & Chang, 2011) besides computers based tests are practical (Bernhardt et al., 2004; Hunt at al., 2007). Recently, the introduction of Dynamic Assessment (Poehner, 2008) has proved that the Sociocultural Theory can have a great impact on the Teaching and testing foreign Languages. This is especially true because the use of connotative elements through visuals, feelings and non-linguistic features that are available through computer based language testing enhances the testee's production.

Current research: The importance of CEFR benchmarking

After more than eight years of research on computer based language testing we have observed the symbiotical benefits of the use of the CEFR in computer based language testing. There are increasing reports that indicate that rating and benchmarking systems have improved significantly in the last years (Lee, Gentile & Kantor, 2010; Attali, Bridgeman & Trapani, 2010; Attali, 2011; Zhang et al., 2012). Automatized systems

[1] http://ec.europa.eu/languages/eslc/docs/en/executive-summary-eslc_en.pdf and http://ec.europa.ec/languages/eslc/docs/en

allow more adequate assessments and permit to place students in the adequate proficiency level. In this sense, computer based testing has benefited from a well-defined framework that permits avoiding biased ratings which occur even among well trained raters. Our current research has also benefit from these developments but since our major concern at the moment is on speaking following the CEFR becomes even more important.

Ubiquitous testing

The most significant problem that computer based language testing has in Spain is the equipment (especially hardware) adequacy in many school. For instance, in some schools in Madrid computers' age can run from one to six year old. This makes difficult using recent software. Additionally, software also changes. While some prefer using free open software (most do), others tend to use the latest versions of windows. In general, the use of tablets is limited and cellular phones are rarely used in educational setting both in general or higher education. Hence, it was important for the research team to work towards a system that allowed to be used in different scenarios. Our main proposals are currently aimed at two major aspects:

- Mobile phone testing: This is probably the most practical approach. It permits a number of exams with limited and inexpensive resources. Interfaces in mobiles have improved dramatically in the last three years. Mobiles have limitations for extensive reading but they can also adequate for other parts of the high school leaving exam (figure 1) including short sentence reading. On the other hand the materials can be provided by the school or just requested.

Figure 1. Alternative prototypes of interfaces of mobile phone based language tests.

- Tablet PC: The use of tablet PC has increased in the last three years. It is very unusual to have classes where tablets are not used commonly in Spain. The major problem with tablets is their wide variety and the different responses that they give depending on their operational system say Android or Windows. At this point, we are currently experimenting on Windows 8. The major drawback is that although it is a versatile for tablets it may not be so much on desktop PC's.

At this point, we still need to see what the requirements for the future high school leaving exam will be but we believe that mobiles may be the most inexpensive solution. In general, they have also proved that student scan easily adapt to their use and would be happy to have them in class.

Only in the last months an alternative design has been suggested which is an external combination of mobile and computer based test. The model presented in figure 2 has its main application duet o the costs incurred due to the design of more sophisticated platforms.

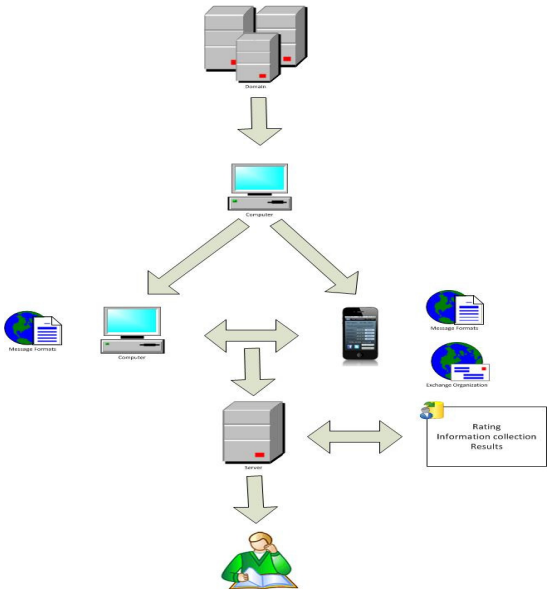


Figure 3. Flux diagram of a combined detachable testing system

The system intends to separate the phone based exam from the regular exam because foreign languages are just one section within the whole exam. In that sense, having a platform that does not need require most the streaming capacity facilitates the data flux for all the other subjects run simultaneously. Although in our opinion this system seems a return in time to much older ones, it has some benefits since the speaking tasks can be done more independently and at different times and spaces. Nevertheless, although this system can be more practical, at this point it is just necessary to add a couple of comments:

Practicality
This test delivery is practical due to its low cost (there is no point in using test delivery means beyond the MECD's limited budget) and limited demands for the new test and is quite accessible even for the students with limited expertise in the use of IT or the media. It is also practical because it permits its delivery in the high school instead of specific places outside the school. Its administration is also relatively easy and the facilitation of the speaking section simplifies the software and hardware requirements in the schools.
Reliability

The delivery system is reliable and permits a consistent data collection because it can be done in comfortable conditions which are optimal for test performance. The principles of modern assessment deal with the testee's emotions and the anxiety reduction. By using familiar IT students have also better possibilities to show their competence and their production in the test would probably be consistent with their performance in the class.

Conclusions: Competence and performance

This paper is first approach to the definition and use of ubiquitous language testing from two different perspectives the use of mobile phones and PC tablets for language testing. Current work on the second type is still in process but depends to a large extent on the limited budget that the new high school leaving exam may have. Technically, both approaches are accessible and will be finished by the end of 2013 but their implementation will depend of external factors rather than on current research. PISA tasks are currently done online in many countries and there seems to be a significant interest in finding ways to deliver the second European Survey on Language Competence. As mentioned before, the use of the CEFR may be more consistent in computer based tests and probably facilitate better assessments in the future.

The researchers would like to express their gratitude to the Ministry of Research and Innovation of Spain (MICINN) for supporting the development and implementation OPENPAU research project (FFI2011-22442) with cofounding with ERDF funds under the 2008-2011 plan. Finally, this paper could not have been possible without the contacts with Matt Poehner and James Lantolf from Penn State University through the Research Stays for Senior Reasearchers grant (with funding from the Spanish Ministry of Education, Culture & Sports, PRX12/00376).

References

AbuSeileek, A. F. (2007). Cooperative vs. individual learning of oral skills in a CALL environment. Computer Assisted Language Learning, 20(5), 493-514.

AbuSeileek, A. F. (2012). The effect of computer-assisted cooperative learning methods and group size on the EFL learners' achievement in communication skills. Computers & Education, 58(1), 231-239.

Alderson J. C. & Wall D. (1993). Does washback exist? Applied Linguistics, 14, 115–129.

Alderson, C. J. (2009). Test review: Test of english as a foreign language[TM]--internet-based test (TOEFL iBT). Language Testing, 26(4), 621-631.

Attali, Y., Bridgeman, B. & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. Journal of Technology, Learning, and Assessment, 10(3), 17-17.

Bernhardt, E. B., Rivera, R. J. & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level language programs. Foreign Language Annals, 37(3), 356-366.

Bridgeman, B., Trapani, C. & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. Applied Measurement in Education, 25(1), 27-40.

Cheng, L. (2003). Looking at the impact of a public examination change on secondary classroom teaching: A hong kong case study. Journal of Classroom Interaction, 38(1), 1-10.

García Laborda, J. (2012). Preliminary Findings of the PAULEX Project: A Proposal for the Internet-based Valencian University Entrance Examination. Journal of language teaching & Research, 3 (2), 250-255.

García Laborda, J. & Fernández Álvarez, M. (2012). Actitudes de los profesores de Bachillerato de Alcalá y Navarra ante la preparación y efecto de la PAU. Revista de Educación, 357, 29-54.

Hew, K. F. & Cheung, W. S. (2012). Students' use of asynchronous voice discussion in a blended-learning environment: A study of two undergraduate classes. Electronic Journal of e-Learning, 10(4), 360-367.

Hirai, A. & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. Language Assessment Quarterly, 6(2), 151-167.

Hunt, M., Neill, S. & Barnes, A. (2007). The use of ICT in the assessment of modern languages: The english context and european viewpoints. Educational Review, 59(2), 195-213.

Jenks, C. J. (2012). Doing being reprehensive: Some interactional features of english as a lingua franca in a chat room. Applied Linguistics, 33(4), 386-405.

Kunioshi, N., Noguchi, J., Hayashi, H. & Tojo, K. (2012). An online support site for preparation of oral presentations in science and engineering. European Journal of Engineering Education, 37(6), 600-608.

Lee, Y., Gentile, C. & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. Applied Linguistics, 31(3), 391-417.

Munoz, A. P. & Alvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. Language Testing, 27(1), 33-49.

Poehner, M. E. (2008). Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development. Berlin: Springer Publishing.

Satar, M. H. & Ozdener, N. (2008). The effects of synchronous CMC on speaking proficiency and anxiety: Text versus voice chat. Modern Language Journal, 92(4), 595-613.

Vetter, A. & Chanier, T. (2006). Supporting oral production for professional purposes in synchronous communication with heterogenous learners. ReCALL, 18(1), 5-23.

Wang, L. & Chang, H. (2011). Improve oral training: The method of innovation assessment on english speaking performance. International Journal of Distance Education Technologies, 9(3), 56-72.

Mario Garcia & Marcin Jaźwiec

CIFESAL, Madrid, Spain
A.H.E., Lodz, Poland

mariogarcia@cifesal.com - mjazwiec@ahe.lodz.pl

Testing Language Competences for an Intended Practical Use

Bio data

L. Mario Garcia is the Director of the International Department of CIFESAL (Centre for Research and Training of Enterprises), a training and consulting centre based in Madrid (Spain), which is certified by the Spanish Education system to provide vocational and occupational training for employment purposes, including foreign languages.L. Mario Garcia graduated in Physics from the University Complutense of Madrid and he has worked in the International field for 18 years in different countries and sectors, in particular research, education and training.At CIFESAL, he has coordinated different transnational projects in EU countries, the Balkan Region, Latin America and China.He has also experience in knowledge transfer, as a certified consultant for foreign trade services to SMEs by the Council of Chambers of Commerce and Industry of Spain and he is member of the teaching staff of the C.E.C.O Foundation in Spain, where he teaches a module on International procurement for an MBA on Multilateral Projects, partially under CLIL modality in English.L. Mario Garcia has published different papers on the topic of the Use of Languages, and Language and Business. (Verdana 10).

Marcin Jaźwiec is the Associate Dean of Philology Faculty at the Humanities Department of the University of Humanities and Economics in Łodź, one of the biggest non-public universities in Poland, offering courses in 11 faculties. Marcin Jaźwiec graduated from the University of Łódź, Faculty of English Philology, in methodology of language teaching and Faculty of Economics and Sociology in the field of Finance and Banking, specializing in corporate finance management. At AHE, he has coordinated bachelor and master degree courses in English Philology, he has been teaching ESP courses to English Philology students such as Business Correspondence, International Business Communication, Business English and International Business Negotiations. He has been author of many articles on CLIL as well as co-author of more than 20 course books for secondary school and university students learning English as a foreign language, FC exam preparation books and customised language courses for the business sector. Marcin Jaźwiec has developed material for many international projects such as Be Multilingual, Promacolt, CCEE, Team Teaching: Transferability and Boundary Zones in Content and Language Integrated Learning (CLIL-AXIS)

Abstract

The paper presents previous conclusions drawn up from a research conducted in the context of the European project Promacolt (www.promacolt.eu), and extends them to the field of language testing based on adapted descriptors of the CEFR.

Promacolt project passed recommendations on how to approach the target users of a foreign language from the perspective of the language course designer/ provider. One of the lessons learnt is the key role played by the course commander, i.e. the organisation which makes possible the actual course (whether a public education organisation with the mission to do so, a private company buying the course for its employees, or a public/ private sponsor, for instance).

According to our findings what a test can actually measure must correlate with the expectations of the language course commander, in particular when it is not a language education organisation itself. Often, there is a gap between one and the other because there is a lack of specification in the first phase of the diagnosis of the intended purpose of the language course. In those cases, we can find a satisfactory performance according to the test, but an inadequate/ incomplete set of competences gained by the learner for the practical purpose expected by the course commander.

The paper illustrates how the spheres of sociolinguistics, psycholinguistics, cultural aspects and pragmatics can be integrated in language testing, by enriching language competence descriptors based on CEFR when language is intended for a practical use.

Short paper

A study conducted in the framework of the EU funded project Promacolt (www.promacolt.eu) taking 12 cases of foreign language training courses, concluded that the targeted use of the foreign language is essential to measure the level of success of the course in practical terms.

In different cases, the expectations of the parties involved are not known by each other, and therefore testing according to existing standards, like the CEFR, may prove not to be sufficient for all parties, even when the results of the test indicates an adequate degree of proficiency achieved by the learners. The level of proficiency measured by the test may not adapt to the use of language expected to be gained by any or several of the parties involved.

Promacolt results encouraged further research in four areas closely related with linguistics, which are considered sources of competences, whose attainment requires testing methods enlarging the purely linguistic indicators. These four areas on which our current research is focusing are: psycholinguistics, sociolinguistics, cultural aspects and pragmatics. In this paper, we just intend to illustrate a few preliminary data.

In the field of psycholinguistics, we are currently focusing on a set of some ten factors, which may contribute to the evaluation of competences in the use of foreign languages, which complement the purely linguistic ones. Some of these factors apply to just one of the parties involved, whereas some other apply transversally to all parties involved, with a different level of influence.

For instance, from the perspective of the learner, his/ her learning style and strategy, and his/ her personality and motivation may substantially condition the way he/ she behaves within the environment of the language test and within the actual environment in which he/ she will use the foreign language. Testing linguistic competences in a safe environment may provide different results than making use of linguistic competences in a hostile environment, or simply under a brand new real situation. Disregarding the psycholinguistic aspects of the learner in a language test has an influence on the potential misperception of his/ her actual competences in the foreign language by other third parties, since the practical use of a foreign language encompasses not only linguistic aspects, but also psychological ones.
The challenges are even higher when we try to analyse the applicability of linguistic tests according to existing standards in order to approach practical uses of the language in social activities involving a group of people working together (or against each other) for a particular purpose. The field of sociolinguistics is extremely relevant to complement language proficiency testing, since most of the standardised testing methods do not incorporate assessment of linguistic abilities in a broad social environment. In the best of the cases, interactive testing incorporates a reduced number of learners, who often are already acquainted with each other.

In our current research, we are focusing on several social factors that we consider relevant to register in order to complement the assessment of purely linguistic tests, in particular when the language course involves a group of learners. Most of these social factors are intended to determine the level of homogeneity of the group, since in our experience and according to the results of our previous research the success of a language course is directly correlated with the homogeneity of the group.

As it happens with psycholinguistics, disregarding the sociolinguistic aspects in a language test has an influence on the potential misperception of the learner's actual competences in the foreign language by other third parties. In the case of foreign language courses for corporate purposes, the impact of social aspects on the use of a foreign language is crucial, since the expectations of the company often rely on conclusive social actions: negotiation of a contract, selection of a product, agreement on a price strategy, etc, in which linguistic competences are deeply interlaced with many other social skills (leadership, communication, negotiation, empathy, etc).

The conclusion of this paper is that CEFR alone is not sufficient to establish testing to measure proficiency on linguistic competences for many different uses of foreign languages. The CEFR is a consistent basis to build upon, in particular to extend and adapt their descriptors to the reality of language practitioners.
Tests based on the CEFR need to be combined with other assessment methods when the intended use of the language is specific, and in particular when the expectations of the organisation commanding the language course go beyond academic ones.

Language testing should be contextualised as much as possible to recreate the practical situations in which the use of the foreign language is intended. For this purpose, it is essential that the language testing incorporate the assessment of the Relevant Learners' Characteristics for the intended Language Use and of the relevant features of the intended Language Use itself.

The assessment of the relevant learners' characteristics and the specification of the intended language use require extending the linguistic competences into metalinguistic competences encompassing: sociolinguistics, psycholinguistics, cultural aspects and pragmatics.

More information about our current research can be found at: www.promacolt.eu/precolt, the webpage of the EU funded project Precolt (Promoting Employment Competences with Language Training).

References

Alltrichter, H. (1993). Teachers investigate their work. Routledge, London and New York.

Allwright, D.& Bailey, K. M. (1991). Focus on the Language Classroom. Cambridge: Cambridge University Press.

Brown, H. D. (1994). Principles of Language Learning and Teaching. Prentice Hall Regents.

Chaudron, C. (1988). Second Language Classrooms. Research on teaching and learning. Cambridge: Cambridge University Press.

Elliot, J. (1991). Action Research for Educational Change. Milton Keynes and Philadelphia: Open University Press.

Ellis, R. (1990). Instructed Second Language Acquisition. Blackwell Publishers.

Ellis, R. (1994). The Study of Second Language Acquisition, Oxford: Oxford University Press.

Ellis, R. (1995). Understanding Second Language Acquisition. Oxford University Press.

Ellis, R. & Wells, G. (1980). Enabling factors in adult-child discourse. First Language 1.

Ely, Ch. M. (1986). An analysis of discomfort, risktaking, sociability, and motivation in the L2 classroom. Language Learning, 36, 1-25.

Eysenck, S. & Chan, J. (1982). A comparative study of personality in adults and children. Hong Kong vs. England. Personality and Individual Differences, 3, 153-160.

Gardner, R. C. (1985). Social Psychology and Second Language Learning: The Role of Attitudes and Motivation. London: Edward Arnold.

Griffiths, R. (1991). Personality and second language learning: theory research and practice.' In Sadtano, E. (ed.). 1991. Language Acquisition and the Second/Foreign Language Classroom. Singapore: SEAMEO Regional Language Centre.

Horn L, R.& Ward G. (Ed.), (2006). The handbook of pragmatics - Malden, MA; Oxford, UK: Blackwell Publishing.

Hudson, R. A. (1996). Sociolinguistics. Cambridge: Cambridge University Press.

Johnson, K. (1992). The relationship between teachers' beliefs and practices during literacy instruction for non-native speakers of English. Journal of Reading Behaviour, 24, 83-108.

Kasper G. & Rose, K. (2002). Pragmatic development in a second language. Malden; Oxford: Blackwell Publishing.

Krashen, S. (1981). Second Language Acquisition and Second Language Learning. Oxford: Pergamon Presss.

Levinson S. (2006). Pragmatics - Cambridge: Cambridge University Press.

Little, D. & Singleton, D. (1990). Cognitive style and learning approach in Duda and Riley (Ed.). 1990. Learning Styles. Nancy, France: University of Nancy.

Reid, J. (1987). The learning style preferences of ESL students. TESOL Quarterly 21.

Richards, J.& Nunan D. (Ed.). (1990). Second Language Teacher Education. Cambridge.

Rose K. & Kasper G. (2001). Pragmatics in Language Teaching. Cambridge University Press.

Skehan, P. (1991). Individual differences in second language learning. Studies in Second Language Acquisition, 13, 275-298.

Spolsky, B. (1990). Conditions for Second Language Learning. Oxford University Press.

Thomas-Malamah, A. (1987). Classroom Interaction. Oxford: Oxford University Press.

Wajnryb, R. (1992). Classroom Observation Tasks. Cambridge University Press.

Wesche, M. B., (1994). Input and interaction in second language acquisition. In Richards, B.J. & Clare, G. (Ed.), 1994. Input and Interaction in Language Acquisition, Cambridge University Press, 219-249.

Williams M. & Burden R. (1997). Psychology for Language Teachers. Cambridge Univeristy Press.

Wright, T. (1987). Roles of Teachers & Learners. Oxford: Oxford University Press.

Yule, G. (1996). The study of language. Cambridge University Press.

Zarate, G., Gohard-Rendenkovic, A., Lussier D.& Penz H. (2004). Cultural Mediation in language learning and teaching. Strasbourg: Council of Europe.

Luke Harding

Lancaster University, Lancaster, United Kingdom

l.harding@lancaster.ac.uk

Investigating the Construct Underlying the CEFR Phonological Control Scale

Bio data

Luke Harding's research interests are in language testing, particularly in the areas of listening assessment, pronunciation and intelligibility, and assessor/rater decision-making. He is also interested in World Englishes and the challenges for language teaching and assessment which are presented by English as an International Language. In the past, he has been involved in several large-scale test development projects, most notably for the Occupational English Test (OET), a test of English for overseas-trained health professionals who wish to practice in Australia and New Zealand. Luke's research has appeared in international peer-reviewed journals, and he is the author of a book, Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test, which was published through Peter Lang in 2011.

Abstract

Pronunciation is a difficult skill to master when learning a second language. It is also a challenging skill to assess, with relatively little research on how pronunciation ability develops, and conflicting views on what the goals of pronunciation instruction should be. The "phonological control" scale of the CEFR represents one attempt at describing pronunciation development. On face-value, the scale presents a useful set of descriptive statements, however, as Harding (2012) has argued, the phonological control scale appears to reflect a view of pronunciation development which conflates increasing ease of understanding with a decreasing level of "foreign accent"; a view which is at odds with the research literature (see Munro, 2008). To date there has been little research on the phonological control scale's validity and usefulness, so any claims that the scale is flawed remain speculative.

The aim of this study was to conduct an empirical investigation of the relationship between ratings of pronunciation ability made using the CEFR phonological control scale and separate measures of comprehensibility (ease of understanding) and strength of accent (the degree to which an accent sounds native-like). 44 non-native speakers of English from a range of first-language backgrounds provided speech samples using a common picture description task. Nine raters then listened to these speech samples and provided ratings on scales designed to measure comprehensibility and strength of accent. The same raters also evaluated the speech samples against the CEFR phonological control performance descriptors, and provided written and oral feedback on their experience of applying these descriptors.

Quantitative and qualitative data will provide evidence from which to draw conclusions about the theory of pronunciation implicit in the phonological control scale. Implications will also be drawn for the usefulness of the scale as a guide both for describing pronunciation development, and for the teaching and assessing pronunciation.

Short paper

Pronunciation assessments come in many different formats (see Harding, 2012). However, language tests which have been designed in the communicative tradition tend to assess pronunciation in a similar way: holistic judgements of pronunciation are made by human raters (who have usually been trained), and pronunciation represents one of several criteria in a broader test of speaking ability. The well-known IELTS format presents a clear example. Pronunciation is one of four criteria assessed during the IELTS Speaking Test, along with fluency and coherence, lexical resource and grammatical range and accuracy. As with the other criteria, pronunciation is assessed along a scale which provides performance descriptors characterising different levels. There is one rater, who is also the interlocutor during the speaking task.

From a theoretical perspective, we can view pronunciation assessment in tests of this kind as taking place within a broader performance model (see McNamara, 1996). From this perspective, we understand that the candidate engages with a task which results in a performance (in this case we can imagine a speaking performance). However the performance does not directly yield a score. Rather, the score is the result of an interaction between the rater and the rating scale which is used to judge the performance. The performance is interpreted through the rating scale, which is, in turn, interpreted by the rater.

It has been recognised that in performance tests, the construct is partly (perhaps mostly) enshrined in scales/criteria (McNamara, 1996; Luoma, 2004). So with respect to pronunciation, the performance descriptors of pronunciation scales indicate what is considered most important in terms of pronunciation ability, and the development which is implied by changes in these descriptions at different levels suggests a theory of pronunciation development. Ideally, scales/criteria should be based on a clear, empirically-based theory of language development. If scales/criteria are not theoretically sound, inferences based on test scores (that is, construct validity) will be weak. However it is sometimes the case that scales will based on intuition, orthodoxy (e.g., an existing syllabus), or on other scales which have already been developed. We must also recognise that, even if we have a very well-formed scale, those using the scale may differ in the way they interpret descriptors, leading to different interpretations of the construct. Differing interpretations will also weaken construct validity because scores will have different meanings. Scales should therefore not only be clearly based on a theory of language development, they should also be interpretable and easy to apply in practice.

One of the problems with many scales of pronunciation has been that it is sometimes unclear what pronunciation construct is being operationalized. In the past the orientation in pronunciation scales was often on "nativeness" (see Levis, 2005), with a trajectory in these scales towards sounding more like a native speaker. More recently, there have been shifts in the wording of scales towards descriptions of intelligibility or comprehensibility. This has largely been the result of research which has shown that a focus on nativeness is neither theoretically justified nor useful. For example, the ongoing work of Munro and Derwing (e.g., Derwing & Munro, 1997) has repeatedly shown that salient features of pronunciation associated with strength of accent may not diminish intelligibility, and that speakers with a strong accent may still be still be very easy to understand. Others, like Jenkins (2000), have argued that native accents might not be most intelligible for all listeners.

However, it is still the case that several influential pronunciation scales continue to conflate dimensions of accent and intelligibility in their descriptors. One example of this is the CEFR phonological control scale (Council of Europe, 2001), which mixes statements concerning strength of foreign accent and ease of understanding in its descriptors. However the CEFR phonological control scale, on face value, presents other problems as well: vague language (e.g., "natural", "clear"), the conflation of lexico-grammatical

knowledge with pronunciation difficulty at the A1 level, and a lack of distinction between level descriptors at the B2 and C1 level (there is no C2 level descriptor). For these reasons, the CEFR phonological control scale represents a useful site for exploring a range of issues related to level descriptors for pronunciation assessment more generally. Specifically, though, as the CEFR phonological control scale may be currently utilised in assessment or teaching contexts, it is important to scrutinise its underlying construct with a view to making suggestions for its improvement.

References

Council of Europe. (2001). Common Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility and comprehensibility: Evidence from four L1s. Studies in Second Language Acquisition, 19, 1-16.

Harding, L. (2013). Pronunciation assessment. Chapelle, C. A. (Ed.), In: The encyclopedia of applied linguistics. Oxford: Blackwell.

Jenkins, J. (2000). The Phonology of English as an International Language. Oxford: Oxford University Press.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. TESOL Quarterly, 39 (3), 369-377.

Luoma, S. (2004). Assessing Speaking. Cambridge: Cambridge University Press.

McNamara, T. (1996). Measuring second language performance. London and New York: Addison Wesley Longman.

Raili Hildén & Marita Härmälä

Finnish National Board of Education, Helsinki, Finland

raili.hilden@oph.fi - marita.harmala@oph.fi

Work in Progress: How Useful is the CEFR in Designing the Follow-Up Assessment of Learning Outcomes in Foreign Languages in the Finnish Basic Education?

Bio data

PhD **Raili Hildén** has some 20 years of expertise as a language teacher and a teacher trainer at the University of Helsinki. She conducts research on language education with a special focus on assessment, proficiency levels, Common European Framework of Reference, and European Language Portfolio. She is a member of the language section of the Matriculation Examination Board responsible for a nation-wide school-leaving examination after upper secondary school. During 2012-2013, she works at the Finnish National Board of Education as a research manager in the national assessment of learning outcomes in foreign languages. The assessment is targeted to the 9th graders in the Finnish Basic Compulsory Education.

PhD **Marita Härmälä** works as a researcher in the Centre for Applied Language Studies at the University of Jyväskylä, Finland. Her main duties include compiling test tasks in in the Finnish Certificate for Language Proficiency, in tests of Swedish, French, and Italian. In addition to this, she also trains item writers and raters. During 2012-2013, she works at the Finnish National Board of Education as a research coordinator in the national assessment of learning outcomes in foreign languages. The assessment is targeted to the 9th graders in the Finnish Basic Compulsory Education.

Abstract

Our presentation will discuss an on-going national assessment on foreign language proficiency administered by the Finnish National Board of Education. Based on the validity theory of criterion-referenced assessment, we explore how scale descriptors serve in designing the tasks to be used for assessing Finnish 9th graders' foreign language proficiency at the final phase of the compulsory basic education. Drawing on the results of related Finnish research projects, we discuss the strengths and challenges of using the CEFR scales and their national application for assessing learning outcomes in general, and for the purposes of the on-going assessment in particular.

Validity of an assessment is a holistic process, starting from defining the test's desirable consequences and ending at conclusions to be made of the results. In the Finnish context, assessments are implemented to evaluate the attainment of the goals set for the language education at school level. Ultimately, the information gathered through the assessments is used for the further development of the National Core Curriculum (2004) for language education. The major instruments for gaining information are language tasks and questionnaires.

Language tasks used in the national assessment have been specified in relation to the domains and goals of the curriculum, which, in turn, reflect the real life language use of the 9th graders. In addition to the appropriateness for the target population, cultural and ethical issues need to be taken into account in the process of task design (McNamara & Roever 2006; Bachman 2010). In the presentation, we demonstrate the task designing

process all the way from choosing relevant proficiency levels and language use situations to compiling the final version of the tests in English, French, German, Russian and Swedish languages.

Short paper

The present project is based on the national legislation about learning outcomes. It is a duty of the Finnish National Board of Education to carry out mappings of how the goals set in the National Core Curriculum are attained in various school subjects at the end of compulsory basic education. The latest evaluation of learning outcomes in foreign languages was conducted in 2001 and in second domestic languages (Swedish and Finnish) in 2008 and 2009. In the present project, learning outcomes are being assessed in five languages (English, Swedish, French, German, and Russian) and two syllabi. In the first place, the study aims at describing, analyzing and explaining learning outcomes in the Finnish foreign language education, and consequently, at informing the design of the next core curricula due 2016. The two main research questions are:

1. To what extent do pupils attain the target levels of language proficiency set in the national core curriculum?
2. What connections are there between learning outcomes and certain background variables (e.g. sex, native tongue, region, socio-economic traits, context of teaching and learning)?

The data are gathered by assessment tasks to pupils and survey questionnaires to pupils, teachers and school leaders. The data include tasks of listening and reading comprehension, speaking and writing, which are designed with regard to the level descriptions provided in the core curricula. The scores from listening and reading comprehension will be correlated with the dependent variables for traditional inferential statistical analyses. Furthermore, modern IRT-based methods (e.g. Bookmark) will be applied for setting the standards of correspondence between performance scores and levels of language proficiency. For speaking and writing, the performances are rated and compared towards the illustrative scales in a straightforward manner.

Features of teaching and learning context, as well as attitudes and conceptions held by pupils, teachers and school leaders are investigated by questionnaires that were informed by of a set of quality indicators published by the Finnish Ministry of Culture and Education. The results of the survey questionnaires are analyzed and reported in relation to these guidelines.

Feedback information on results will be delivered to municipal authorities and school leaders. The final outcome of the assessment is compared to the goals of the national core curricula to detect strenghts and weaknessess. Based on the results, state of art of the Finnish language education will be discussed and a set of recommendations for improvement will be published to inform school administrators and teachers. Further decisions on how to use this information are made locally.

Degree of difficulty of the levels
How can we make sure that our examinations are measuring at the CEFR levels we claim they are? What evidence do we have to support our claims?

Following measures are taken to acertain the quality (primarily in terms of level correspondence) of the tasks and items:

1. Since the assessment tasks are targeted to measure certain levels given in the National Core Curriculum for each syllabus, the item writers were familiarised with the level system and requested to produce tasks to match the following distribution:

50% of the tasks onto the target level assigned for the syllabus, 25% onto the adjacent levels above and below.
The descriptions address a range of communicative and linguistic features at each level. In long syllabus English, for instance, Finnish school leavers are expected to cope with the following kind of reading comprehension in order to aquire an average grade of good mastery at the end of compulsory basic education. The description corresponds to the lower band of the CEFR level B1.

- Can read texts of a few pages (newspaper articles, brochures, instructions, plain language authentic prose, and letter from a pen-friend) even unprepared
- Can follow the main ideas, key words and important details in a text of a couple of pages dealing with familiar background knowledge and experience.
- Understanding unfamiliar topics and details can cause problems.

In line with the discourse types and processes mentioned above, texts and tasks were selected for the assessment by the group of item writers.

For speaking tasks the following specification was consulted:

- Can give a short linear description on a personally relevant, familiar topic. Can communicate in routine exchanges and short conversations on personally relevant topics. May need some assistance or avoid certain topics.
- The speech is reasonably fluent, but various kinds of pauses are still very evident.
- Pronunciation is intelligible despite a noticeable foreign accent and phonological mistakes.
- Shows a reasonably good command of everyday vocabulary and a limited number of idioms. Can use a variety of simple structures and some complex ones.
- Makes a lot of basic mistakes (e.g. verb tenses) in more extended spoken production and these may occasionally cause intelligibility problems.

Despite the emergently straightforward formulations, the use of scale descriptions involves certain unavoidable problems. In case of receptive skills, text length is a complicated feature. Extensive texts can be simple, and on the other hand, a short text can be extremely demanding. Nor is the concepts of familiarity as clear-cut as it may seem at first sight, since themes and situations are of varying familiarity to different pupils. Regarding productive skills, features of grammatical range and accuracy are frequently resorted to, but even there, great care should be taken in defining for example "everyday vocabulary" and "basic mistakes" in a language.

2. As a part of the assessment design procedure, the item writers judged in advance the intended level of each MC and open-ended test item for reading and listening as well as for speaking and writing.

3. Based on the test-taker data from pre-tests, a tentative standard setting procedure was conducted on receptive skills (listening and reading) in English and Swedish.

1. The items chosen for the final test form were arranged in a booklet in the order of difficulty according to the discrimination index obtained for each item in the pre-tests.
2. The panelists representing each language were asked to identify cut scores and place a bookmark in the booklet.
3. At the final stage of the standard setting procedure, the panelist judgements are to be compared with empirical data from the final outcomes of the nation-wide assessment.

4. Test-taker results are reported by means of proficiency levels. The outcome is compared to the assumption made for the specific syllabus in the core curriculum.

References

Bachman, L. & Palmer, A. (2010). Language testing in practice. Oxford: Oxford University Press.

Cizek, G. J. & Bunch, M. B. (2007). Standard Setting. A Guide to Establishing and Evaluating Standards on Tests. Sage Publications, Inc.

European Council. (2001). The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Hildén, R. & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen & V. Kohonen (Ed.), Foreign languages and multicultural perspectives in the European context, 291–300. Reihe: Dichtung – Wahrheit – Sprache Bd. 7. Münster: LIT Verlag.

McNamara, T. & Roever, C. (2006). Language Testing: The Social Dimension. Hoboken, NJ: Wiley-Blackwell.

The Finnish National Board of Education. (2004). National Core Curriculum for Basic Education.
http://www.oph.fi/english/sources_of_information/core_curricula_and_qualification_requirements/basic_education

Ari Huhta, Riikka Ullakonoja, Lea Nieminen & Eeva-Leena Haapakangas

University of Jyväskylä, Jyväskylä, Finland

ari.huhta@jyu.fi - riikka.ullakonoja@jyu.fi - lea.s.m.nieminen@jyu.fi - eeva-leena.haapakangas@jyu.fi

The Use of the CEFR in Diagnosis

Bio data

Ari Huhta is a professor of language assessment at the Centre for Applied Language Studies at the University of Jyväskylä, Finland. He has participated in several test development and research projects, many of which have concerned using the CEFR and/or linking assessments to it (e.g. DIALANG and currently the DIALUKI project).

Lea Nieminen, PhD, is currently a researcher in the DIALUKI project. She specializes in Finnish as L1 and L2 and in the development of morphology in learners' language.

Riikka Ullakonoja, PhD, is currently a researcher in the DIALUKI project. She specializes in Russian as a foreign language, phonological development and motivation.

Eeva-Leena Haapakangas, MA, is currently a research assistant in the DIALUKI project. She specializes in the Russian language.

Abstract

This presentation reports on how the CEFR has been utilized in a research project focusing on the diagnosis of second or foreign language (SFL) proficiency. The talk first gives an account of the role of the CEFR in the design of the study and, secondly, reports on findings that shed light on the relationship between certain areas of linguistic knowledge and reading and writing at different CEFR levels.

The project is an international 4-year (2010-2013) study into the diagnosis of SFL reading and writing. It seeks to identify the cognitive, affective and linguistic features which predict a learner's strengths and weaknesses in those areas by studying several hundred SFL learners both cross-sectionally and longitudinally. The project aims at deepening our understanding of L2 development and of the factors that affect it, and will formulate hypotheses for further work on SFL diagnosis. The project relates to the activities of the European SLATE (Second Language Acquisition and Language Testing in Europe) research network (see www.slate.eu.org)

We give an account of the ways in which the CEFR influenced the design of the study. First, we describe the selection of CEFR-related reading tests such as DIALANG and the Pearson Test of English General for data-gathering purposes. Secondly, we report on the success of applying CEFR-related scales for rating writing performances in SFL and L1. In this way, it was possible to arrive at CEFR-referenced estimates of the informants' reading and writing ability in two languages.

We also report on findings about the relationship between CEFR levels and linguistic aspects of performance. For example, we provide a characterization of the CEFR levels in reading and writing English in terms of a vocabulary profile. Finally, we reflect on the usefulness and limitations of the CEFR for SFL diagnosis.

Short paper

The use of the CEFR in diagnosis

This paper reports on an international 4-year (2010-2013) study into the diagnosis of SFL reading and writing. The project seeks to identify the cognitive, affective and linguistic features which predict a learner's strengths and weaknesses in those areas by studying several hundred SFL learners both cross-sectionally and longitudinally. The study aims at deepening our understanding of L2 development and of the factors that affect it, and will formulate hypotheses for further work on SFL diagnosis. The project relates to the activities of the European SLATE (Second Language Acquisition and Language Testing in Europe) research network (see www.slate.eu.org).

We give an account of the ways in which the CEFR influenced the design of the study. First, we describe the selection of CEFR-related reading tests such as DIALANG and the Pearson Test of English General for data-gathering purposes. Secondly, we report on the success of applying CEFR-related scales for rating writing performances in SFL and L1. In this way, it was possible to arrive at CEFR-referenced estimates of the informants' reading and writing ability in two languages. We will also mention some results concerning the relationship between vocabulary and CEFR levels.

Although the CEFR has been very influential in language testing and many tests have been linked with the CEFR levels, it is not easy to get access to carefully developed, validated and CEFR-referenced language tests in order to use them as data gathering instruments in an applied linguistics study. Since DIALUKI members represented core partners in the project that developed DIALANG in the early 2000s, we had access to all DIALANG tests, i.e., to tests that had been aligned with the CEFR. In fact, DIALANG had been the first large-scale language testing system to have been designed by making use of the content of the CEFR and by linking its test scores with the CEFR scale. Thus, the project was in the fortunate position to be able to use DIALANG reading tests of English and Finnish as measures of SFL reading comprehension. DIALANG converts its test scores to CEFR levels on the basis of an algorithm based on the analysis of the standard setting data gathering during the DIALANG project.

The DIALUKI research project faced a number of challenges, however. First, we had to design our own online test delivery system to administer DIALANG tests as part of the study, as it was not possible, for practical reasons, to extract item-level data from the operational DIALANG system. Second, DIALANG tests are not suitable for young learners, so we had to use other, non-CEFR-related reading tests to measure the SFL reading skills of our youngest target group, the primary school students involved in the study.

To measure the construct of SFL reading more comprehensively, we also used a selection of Pearson Test of English General reading items. Thanks to other, related research cooperation between DIALUKI members and Pearson, we could use operational PTE General reading items that were targeting specific CEFR levels. However, since we did not use a full PTE reading test, converting the test score based on the PTE items to the CEFR level is not straightforward and has not yet been attempted in the project.

When it comes to assessing writing in SFL, the situation is somewhat different. Rating SFL writing against the CEFR levels can in principle be done in at least three different ways: 1) use tasks that represent specific CEFR levels, 2) use a rating scale taken from the CEFR or specifically designed to be related in some systematic way to the CEFR scale, or 3) use a rating scale that is afterwards linked with the CEFR levels in some systematic way. Since the project members had already experience with using the second approach in another project (the CEFLING study of writing in SFL), that approach was adopted for DIALUKI (Alanen, Huhta & Tarnanen 2010).

The Study 1 writing performances in both L1 and SFL were rated by teams of raters using the 6-point CEFR scale. The descriptors of the levels were compiled from six different writing scales found in the CEFR that were considered to be usable and relevant to the writing tasks used in the study. Each performance was rated by two or three trained raters and the rating data were analysed with the Multi-faced Rasch Measurement software Facets (see Alanen, Huhta & Tarnanen 2010, for more information about a similar study). In later stages of the project, writing performances were also used against a CEFR-referenced scale that was developed for the Finnish national curriculum. The scale has more levels than the CEFR (each level is divided into at least two sublevels) and the content is modified to suit the rating of younger learners' performances better. The results of the analyses indicate that both the CEFR and the national curriculum scales worked as rating scales and, thus, there is some confidence in our placement of learners' scripts on the CEFR scales.

In another part of the study, we analysed the relationship between the results of the Vocabulary Levels Test (for English) (see Schmitt, Schmitt & Clapham 2001) and reading and writing in English. Of particular interest was to find out if there was a systematic relationship between the results of the VLT and the CEFR results assigned to the learners based on their performance in the writing tasks (that were rated directly onto the CEFR levels) and in the DIALANG English test of reading (that converts scores onto CEFR levels on the basis of an algorithm derived from the results of standard setting the test items used in that test). We found, first, that vocabulary test results and reading and writing correlate quite strongly, which is in line with a lot of previous studies. More interestingly, however, we were able to create a vocabulary profile for each CEFR level that indicates the proportion of English words at each frequency band (e.g. 2000 word level, 3000 word level, and so on) that a typical reader or writer at CEFR levels A1 to B2 has -- at least in our data.

In summary, in the DIALUKI project, the CEFR has turned out to be a useful overall framework to conceptualise and define in practice different levels of functional language proficiency. The CEFR does not in itself contain detailed enough information about factors that we have been interested in such as the specific linguistic features of the languages that we have studies or the cognitive aspects of language knowledge and use (e.g. working memory, phonological processing, access to words). However, the fact that the CEFR provides definitions of levels of proficiency that have turned out to be useful for rating purposes (rating of writing performances) and also for defining levels of comprehension (via reading tests that were linked with the CEFR), has been important for the project. It provides a common frame of reference against which to interpret some of the more detailed, diagnostic findings about the strengths and weaknesses in leaners' foreign and second language proficiency.

References

Alanen, R., Huhta, A. & Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. In I. Bartning, M. Martin & I. Vedder (Ed.) Communicative proficiency and linguistic development: intersections between SLA and language testing research. EUROSLA Monograph Series, 1. Pp. 21-56.
http://eurosla.org/monographs/EM01/EM01home.html

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Level Test. Language Testing, 18, 55-88.

Ben Knight

Cambridge University Press, Cambridge, United Kingdom

bknight@cambridge.org

The English Profile Project: Researching what the CEFR Means in Terms of Specific English Linguistic Knowledge

Bio data

Ben Knight joined Cambridge University Press two years ago and is now their Director of Language Research and Consultancy. One of his key interests is to develop the way CUP invests in research to improve the effectiveness of its learning materials. He also works closely with Cambridge English Language Assessment to enable the two university departments to offer an integrated service to teachers and organisations around the world. Ben has worked for the British Council, International House, Cambridge English Language Assessment (then UCLES) and City & Guilds. He holds an M.Sc. in Applied Linguistics.

Abstract

The English Profile Project has supported research projects into what the CEFR levels mean explicitly for English, the Reference Level Descriptions, particularly in terms of grammar and vocabulary. This presentation will share some of those findings, particularly those from research based on the Cambridge Learner Corpus. It will also share some of our work into linking English Profile to curriculum development. Although this presentation relates to research, it is more focused on its significance for teaching and learning than a formal research paper.

Short paper

English Profile is a long-term, collaborative programme of interdisciplinary research to produce Reference Level Descriptions for English to accompany the CEFR (Common European Framework of Reference; Council of Europe 2001). These descriptions cover what learners know and can do in English at each of the six CEFR levels. English Profile is registered with the Council of Europe and is currently managed by a core group of collaborators at the University of Cambridge.

The CEFR has been very useful as a planning tool that provides a 'common language' for describing objectives, methods and assessment in language teaching, as put into practice in diverse contexts for many different languages. However, it is insufficient on its own to provide the detail needed for professionals to make important decisions about teaching and testing. Importantly the CEFR is neutral with respect to the language being learned. This means that the users have to decide what actually gets taught or assessed in terms of the linguistic features of a specific language at each of the common reference levels. This is where the development of Reference Level Descriptors (RLDs) for each language is important. English Profile aims to deliver those RLDs for English.

An innovative feature of English Profile, distinguishing it from previous work in this field, is that research is based on electronic corpora of learner data, including the largest analysed corpus of language learner output in the world: the Cambridge Learner Corpus. This research approach is producing results which can be empirically measured and which are not predictable from current language learning theories alone. An important future

focus is also the impact of different first languages and learning contexts and the effects of language transfer on learning at the different CEFR levels (A1 to C2).

The Cambridge Learner Corpus (CLC) has been at the centre of this work to-date. This is a large corpus of learner language which consists of more than 200,000 written Cambridge exam scripts and currently contains over 50 million words. It also has some important features which are not found in other L2 learner corpora:

- it is larger than most learner corpora (and continues to grow by around 3 million words a year)
- the scripts have been systematically categorised by their CEFR level according to reliable information captured during the examination process
- a large amount of information is stored about the learners, including their L1
- parts of the corpus have been coded for errors
- the corpus has been tagged and parsed using computer programmes developed by computational linguists at the University of Cambridge Computer Laboratory.

The information about the learners allows researchers to compare different L1 learners with respect to the English that they produce; around 100 L1s are represented (from 200 different countries), over 20 of which with samples large enough for quantitative analyses to be carried out. The error coding and the parsing means that, in addition to lexical analysis, sophisticated kinds of grammatical analysis are also possible.

A wide range of projects have been linked to the English Profile programme, each one shedding light on different aspects of proficiency in English at different CEFR levels. This paper focuses on two projects which Cambridge University Press has played a leading role: the English Vocabulary Profile and the English Grammar Profile.

The English Vocabulary Profile (EVP) project has set out to describe the vocabulary that learners of English typically master at each CEFR level. It covers words, phrases, phrasal verbs and idioms. It is unique in its approach to polysemy – recognising that learners master different meanings of a word at different stages of their learning as they progress through the CEFR. The levels at which learners master each meaning were determined by a team of lexicographers, drawing largely on the Cambridge Learner Corpus, but also drawing on other sources of data: the corpus-based Cambridge dictionary data that gives information on native speaker usage and frequencies of use for each word/phrase (and each meaning within those), major course-books, commonly used wordlists, exam item writer guidelines, etc. Although the principle of English Profile is to be corpus-driven, we recognise the limitations of the corpus data we currently use and so have cross-checked it against these other sources.

The English Vocabulary Profile research led to the development of the EVP Online Resource – a free online searchable database of our findings. This enables materials writers, test item writers, curriculum developers and teachers to check the typical level of words, phrases and meanings, produce lists of words for particular topics at certain levels, see how different meanings of words fit across CEFR levels, compare American and British English, etc.

The English Grammar Profile project is also under way, and aims to describe how learners of English typically master features of English grammar at each CEFR level. This has been analysed from both a formal and a usage point of view. So, for example, in analysing mastery of the modal verb 'can', we see the affirmative, negative and question forms mastered at A1, with negative questions ('Can't you...') at B1 and the perfect form ('It can't have been..') at C1 level. In a parallel with lexical polysemy, it is the different uses or meanings of the grammatical feature which can be particularly interesting. With the same example of the modal verb 'can', we see it being used to express possibility, ability and requests at A1, permission at A2, general truths/tendencies at B1,

guesses/predictions at B2, emphasis at C1, and reflection at C2. This research is giving us clear data about the way learners develop proficiency in grammar, not just by mastering the forms but by mastering the full range of their functional usage as well.

The next key English Profile project that Cambridge will be focusing on will be on spoken English. We are building up our corpus of learner spoken English, structured with data on CEFR levels and L1, to enable us to address two key areas of research: i) how do our findings on vocabulary and grammar change when the analysis is based on spoken data? and ii) how do other aspects of speaking – pronunciation, fluency, interactional aspects (turn-taking, showing engagement, etc) typically develop across the CEFR levels?

Cambridge University Press has been using English Profile research to reshape its approach to curriculum development within its courses. English Profile is explicitly non-prescriptive; the outcomes of the research do not tell us what needs to be taught at a particular level. However, it does provide evidence-based insights into what should be considered at each level for a course. This can be done with different degrees of focus: for example, we can ask for a report on the way modals are typically mastered across levels A1-C1 in order to plan out how modal verbs are treated across each level of a 5-level course. We can ask more specific questions such as 'what are the adjectives used for describing things at B1 and B2 levels?' Or an author team might use the English Vocabulary Profile to check whether a particular use of a word is really suitable for an A2 course.

Many other factors have to be taken into account in the finalisation of a curriculum or course – the particular target studentship, their linguistic and educational background and their learning objectives, how heterogeneous they are, topics of interest, and a certain degree of logical coherence (e.g. the vocabulary for close family relationships is most helpfully presented as a single set). English Profile does not set up to specify a single size-fits-all syllabus, but provides the resources to make more evidence-based decisions that are likely to improve the effectiveness of the course programme by being better targeted at learners' needs at each level.

English Profile was set up as a collaboration between the University of Cambridge, (incl two of its departments, Cambridge English Language Assessment and Cambridge University Press), the University of Bedfordshire, the British Council and English UK. But a key feature of the programme has been the development of the English Profile Network – a wide range of academics, government advisors and educationalists that are contributing to its objectives, by providing their own data or research, or simply by participating in English Profile workshops and seminars. Individuals and institutions around the world are welcome to join the Network – see www.englishprofile.org for more details and a lot more information on the programme and its outputs.

The English Profile Programme is directly addressing the question of competence and performance: how can we give evidence-based answers to the issue of what linguistic knowledge and skills should be taught at each CEFR level? The programme is not approaching this question by starting with the Can Do statements, but by analysing the linguistic skills of learners placed at specific CEFR levels through validated assessments. A future step in the programme, however, is expected to include tagging corpus data from a functional/notional perspective, to analyse realisations of Can Do statements and link to the current analysis of linguistic competence.

Benjamin Kremmel & Franz Holzknecht

University of Innsbruck, Innsbruck, Austria

benjamin.kremmel@uibk.ac.at - fholzknecht@gmail.com

Strengths and Weaknesses of the CEFR in Guiding Test Task Design: What the Can Do's Can Do and What They Can't (yet)

Bio data

Benjamin Kremmel holds a degree in English Language Teaching, Philosophy and Psychology from the University of Innsbruck, Austria. He is professionally involved in the development of the new standardized Austrian school-leaving examination for the modern foreign languages and is a member of the newly established research group "Centre for Language Education and Assessment Research (CLEAR)" at the University of Innsbruck, Austria.

Franz Holzknecht holds a degree in English Language Teaching and Sports Education from the University of Innsbruck, Austria. He was a member of the University of Innsbruck's team that develops the new standardized Austrian school-leaving examination for the modern foreign languages from 2008 to 2012. He is currently living in Boston, MA.

Both authors have submitted their dissertations for Lancaster University's Language Testing MA program in December 2012.

Abstract

This practice-related paper will report on opportunities and limitations the CEFR has provided in setting up a national school-leaving exam for modern languages in a European country. While the framework has been particularly helpful in drawing up competence-based test specifications for the traditional language skills, test developers are often faced with shortcomings of the CEFR when it comes to the implementation and operationalization of the level descriptors on a practical level. Although the framework states that task support (e.g. instructions), text characteristics (e.g. text type, discourse structure, presentation, length, relevance and linguistic complexity) and the type of response required can affect the difficulty of comprehension tasks (Council of Europe, 2001, p. 164f), the CEFR lacks specificity when it comes to translating these features into levels on the illustrative scales. On what basis is it then that test developers decide whether a text for a reading comprehension task is at B2 level, whether the response format used is suitable for A2, or whether a writing task will elicit a B2 performance rather than a B1 performance? Such challenges become particularly prevalent when it comes to the design of tasks for the productive skills and to ensuring fairness and equal difficulty for all candidates taking parallel versions of a test. The paper will present advantages and drawbacks experienced throughout five years of working with the CEFR in large-scale assessment and will suggest that the framework can and should be refined for assessment purposes.

Short paper

The CEFR has become the basis of many curricula and school-leaving exams in Europe. In the practice of setting up such high-stakes tests, however, it has shown both strengths and weaknesses when it comes to the operationalization of descriptors in the

phases of task design. A team of language testers has experienced these assets and shortfalls of the CEFR over the last five years of developing a CEFR-based standardized school-leaving exam for the modern foreign languages in Austria. Selected insights from this process are exemplified on the skills of reading and writing in the following.

The framework has been particularly helpful in drawing up competence-based test specifications for the language skills as a basis for task design. For reading at B2 level, for example, descriptors such as "[c]an understand articles and reports concerned with contemporary problems in which the writers adopt particular stances and viewpoints" (Council of Europe, 2001, p. 70) or "[c]an read correspondence relating to his/her field of interest and readily grasp the essential meaning" (Council of Europe, 2001, p. 69) have proven to be easily operationalizable in test situations. Other descriptors, or rather aspects of descriptors, have turned out to pose challenges for test developers, at least in the Austrian context. In operationalizing the descriptor "[c]an read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively" (Council of Europe, 2001, p. 69), for instance, task designers face problems of practicality such as how to measure "independence" or how to set time limits for reading tasks. Similarly, the descriptor "[c]an scan quickly through long and complex texts […]" features a shortcoming of the CEFR for item writers as terminology is being used vaguely and not defined in detail. What the CEFR considers "long" is as opaque as is the question whether this descriptor refers to linguistic or thematic complexity, or both.

The question remains then, how item writers deal with such vaguenesses of the CEFR when they design comprehension tasks. Thus, although the framework suggests that task support (e.g. instructions), text characteristics (e.g. text type, discourse structure, presentation, length, relevance and linguistic complexity) and the type of response required can affect the difficulty of comprehension tasks (Council of Europe, 2001, p. 164f), it lacks specificity when it comes to translating these features into levels on the illustrative scales, as has already been pointed out by the Dutch Construct Project team (Alderson et al., 2004; 2006). In an attempt to overcome this problem, reading test booklets were analysed by means of the Dutch CEFR Construct Grid, suggested for linking exams to the CEFR by the Manual (Council of Europe, 2009). It was found, however, that this grid also lacked key specifications and instructions, making the results it generated only partly useful to inform further comprehension task design.

When it comes to designing writing tasks, the relationship between task type, task complexity and L2 writing performance is not yet clear (Alanen et al. 2012; Hamp-Lyons & Mathias, 1994; Kuiken & Vedder, 2008). Taylor and Weigle (2012) state that it is questionable to what degree a writing task can be assigned to a particular level, or if it can be assigned to a level at all. According to them, it is usually the performance, and not the task, which determines whether a candidate is at a certain level. The CEFR agrees with this observation to some extent, as it states that "tasks may be designed so that the same input may be available to all learners but different outcomes may be envisaged quantitatively (amount of information required) or qualitatively (standard of performance expected)" (Council of Europe, 2001, p. 159). It is not clear to test developers how these features can be altered to accommodate for the difficulty of the different levels. It seems that one way to alter the difficulty of writing tasks might be by asking test takers to showcase different degrees of functional competence. For example, a writing task asking test takers to describe an event might produce linguistically less complex language than a writing task asking to argue a viewpoint. Macrofunctions, as defined by the CEFR, "are categories for the functional use of spoken discourse or written text consisting of a (sometimes extended) sequence of sentences" (Council of Europe, 2001, p. 126). Although the framework lists a number of macrofunctions, such as description or persuasion, the list is not extensive. Moreover, the different functions are not assigned to specific levels. It would be helpful for test developers to have guidance on which of these functions test takers should be able to perform at different levels, and to what degree. Corpus linguistic studies such as CEFLING (Alanen et al., 2012) might be

able to shed some light on these questions, which is why a Learner Corpus compiled of written performances of Austrian students has been set up to inform further task design and address the often accused imperfection of the CEFR regarding its lack of foundation on empirical evidence from L2 learner data (Hulstijn, 2007).

Another area where there might be room for refinement of the CEFR in terms of guidance for writing task design concerns text types. A strength of the CEFR is that it links specific text types to different levels, for example "postcard" to A1 or "essay" to B2 and C1 (Council of Europe, 2001, p. 26f). However, many of the text types listed appear in more than one level and the number of different text types does not seem to be exhaustive. Moreover, it might be time to rethink the inclusion of text types such as "letter", seeing that most correspondence is done by email nowadays. Through needs analyses of the target language use situations of learners on which text types are produced in real life the list of text types could be extended and refined. In addition, it was felt in the Austrian exam reform that there is a strong need among testers and teachers to have access to clearly defined text type definitions, both for task design and standardized assessment purposes. Such a document has been established for the Austrian context. A similar endeavour might be necessary for a future refined version of the CEFR, as interpretations on what it entails to write an essay, for example, might vary between different countries.

A particular strength of the CEFR, for task design for both the receptive and the productive skills, is that it provides a list of potential topics that are suitable for language assessment in Table 5 (Council of Europe, 2001, p. 48f). Although this is usually perceived to be a useful guideline for task designers, practical experience shows us that, more often than not, item writers choose topics of interest first and then attempt to accommodate it to the categories of Table 5, thus using these more as a lifeline rather than a guideline. A similar behaviour can be observed with the potential sources and discourse types listed in the CEFR, which might again be due to the lacking relation of these task features to concrete competence levels, or due to the fact that these types have undergone significant changes themselves. A framework of 2013 might need updating with, for instance, aspects of writing or reading impacted by developments in technology and new media.

Daily practice in item moderation shows that item writers often check input material for comprehension tasks with text analysis tools such as Lextutor, CohMetrix or corpus analyses with the help of the COCA or the BNC. However, none of these indices have been mapped against the CEFR levels yet, indicating a dire need for research and (language-) specific supplementation or adaptation of individual descriptors. In doing so, and in complementing CEFR descriptors with corpus data, another advantage of the CEFR could come into effect, namely that it understands itself as flexible, open and dynamic, i.e. adaptable and capable of refinement "in response to experience in its use" (Council of Europe, 2001, p. 8). Language testing research, by gathering experience in operationalizing the CEFR and by generating valid and reliable L2 user data, can thereby contribute to constructing and fine-tuning the "poles underneath" (Hulstijn, 2007, p. 666) the CEFR building to ensure this Common European Framework stands on a "proper foundation" (Hulstijn, 2007, p. 666).

References

Alanen, R., Huhta, A., Martin, M., Tarnanen, M., Mäntylä, K., Kalaja, P. & Palviainen, Å. (2012). Designing and assessing L2 writing tasks across CEFR proficiency levels. In Torrance, M., Alamargot, D., Castelló, M., Ganier, F., Kruse, O., Mangen, A., Tolchinsky, L. & Van Waes, L. (Eds). Learning to write effectively: current trends in European research. Bingley: Emerald Group Publishing Limited. Retrieved February, 2013, from http://eurosla.org/monographs/EM01/21-56Alanen_et_al.pdf

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). The development of specifications for item development and classification within The

Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening: final report of The Dutch CEFR Construct Project. Project Report. Lancaster University, Lancaster, UK. Retrieved June 12, 2010, from http://eprints.lancs.ac.uk/44/

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: the experience of the Dutch CEFR Construct Project. Language Assessment Quarterly, 3 (1), 3-30.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Retrieved June 12, 2010, from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Council of Europe. (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): a manual. Retrieved June 12, 2010, from http://www.coe.int/t/dg4/linguistic/Source/Manual%20Revision%20-%20proofread%20-%20FINAL.pdf

Hamp-Lyons, L. & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. Journal of Second Language Writing, 3, 49–68.

Hulstijn, J. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. The Modern Language Journal, 91, 663-667.

Kuiken, F. & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. Journal of Second Language Writing, 17, 48–60.

Taylor, L. & Weigle, S. C. (2012). Assessing writing. Preconference workshop at the 2012 EALTA conference, Innsbruck.

Folkert Kuiken & Ineke Vedder

University of Amsterdam, Amsterdam, The Netherlands

f.kuiken@uva.nl - i.vedder@uva.nl

Functional Adequacy as a Fundamental Component of L2 Proficiency

Bio data

Folkert kuiken is Professor of Dutch as a Second Language at the University of Amsterdam, where he coordinates the Dual Master of Dutch as a second language. His research interests include the effect of task complexity and interaction on SLA, Focus on Form, and the relationship between linguistic complexity and communicative adequacy.

Ineke Vedder is Senior researcher and Head of Education at the University of Amstredam. Her research interests include instructed SLA, particularly Italian as a second language, pragmalinguistics, the influence of task complexity and interaction on L2 performance, and the relationship between linguistic complexity and communicative adequacy in L2 writing.

Abstract

In recent studies in SLA often general measures for assessing linguistic performance, such as complexity, accuracy and fluency (are employed (Housen, Kuiken & Vedder 2012). Few studies, however, report on the functional adequacy of the L2 output, considered as being crucial by some authors (De Jong et al. 2012; Kuiken, Vedder & Gilabert 2010, Pallotti 2009). Until now, there has been no unanimity, as to how functional adequacy is to be defined or assessed and by which features it is determined (Iwashita et al. 2008). While functional adequacy is sometimes interpreted as socio-pragmatic appropriateness (McNamara & Roever 2007), in other cases it is operationalized as communicative effectiveness (Upshur & Turner 1995) or successful task completion (De Jong et al. 2012).

The primary focus of the present paper is on the assessment of L2 writing. It is argued that the assessment of linguistic performance in L2 is not possible without taking into account the functional dimension of L2 production, as defined in the CEFR. The analysis is based on the written output of 32 learners of Dutch and 39 learners of Italian at CEFR level B1, who were submitted to two argumentative writing tasks. Functional adequacy was rated by experienced raters on a 6-point Likert scale. During a subsequent panel discussion, the raters were asked to verbalize the reasons underlying their decisions to assign a text to a particular rating level.

In the paper the following questions will be discussed (1) How can functional adequacy as a construct be defined and measured? (2) What are the features of functional adequacy that raters consider to be crucial? (3) How can differences between low ptoficient and high proficient L2 learners be described in terms of functional adequacy?

Short paper

The notion of language proficiency presented in the Common European Framework (CEFR) rests on two pillars, as has been pointed out in several studies (Hulstijn, 2007).

Language proficiency is defined both functionally ('can-do statements'), describing the number of domains, functions and roles language users can deal with in the L2 (what), and in terms of the quality of language proficiency, e.g. the degree to which language use is effective, precise and efficient. Whereas the majority of research conducted so far has been concerned with the can-do-statements and the functional scales of the CEFR (Little, 2007), fewer studies have focused on the linguistic dimension, particularly regarding the question of whether it is possible for L2 learners to be situated at different linguistic scales and levels (for instance the B1 level for vocabulary range, and the A2 level for grammatical accuracy), or the specific ways in which L2 proficiency develops in different European languages. Moreover, the CEFR doesn't indicate, for a given target language, which particular developmental features can be identified as being characteristic for a given scale level (Alderson, 2007). The relationship between language proficiency and language acquisition and the overall development of L2 proficiency (in terms of syntactic complexity, lexical diversity, fluency and accuracy) and the way in which they interact, is thus still unclear (Hulstijn, 2007, 2010).

The relationship between the functional descriptor scales of the CEFR on the one hand and the linguistic scales on the other hand has not been addressed much in the literature either. One of the few studies which have investigated the relationship between the functional and the linguistic dimension of L2 performance is the so called WISP study ('What Is Speaking Proficiency'; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2007, 2012). In the WISP study the oral performance of 208 L2 speakers and 59 native speakers of Dutch was examined both in terms of communicative success and in linguistic terms, concerning the mastery of a number of linguistic skills, such as fluency (i.e. breakdown fluency, speed fluency and repair fluency), syntactic complexity, and vocabulary control. The main question in this type of research is to what extent it may be expected that L2 learners who are situated at the B2 level of the functional descriptor scales of the CEFR have also attained the B2 level with regard to their linguistic performance. In other words, the issue at stake is if and how the communicative adequacy of L2 performance ('getting the message through') is related to the syntactic complexity, lexical variation, fluency, and accuracy of the output.

In recent studies on L2 performance generally measures for assessing the linguistic complexity, accuracy and fluency (CAF) are employed (Housen & Kuiken 2009; Housen, Kuiken & Vedder 2012; Wolfe-Quintero, Inagaki & Kim 1998). Few studies, however, report on the communicative success and functional adequacy of the L2 output. This is in contrast with how these are treated in language teaching and testing, where an effort is made to independently assess functional adequacy on the one hand and linguistic complexity and accuracy on the other hand (Pallotti 2009). However, for a clear understanding and interpretation of L2 proficiency functional adequacy needs to be taken into account, next to syntactic complexity, lexical diversity, accuracy and fluency.

At the moment a coherent and clear-cut definition and operationalization of functional (or communicative) adequacy does not exist (Housen, Kuiken & Vedder 2012, Kuiken, Vedder & Gilabert 2010; Pallotti 2009). While functional adequacy is sometimes interpreted as socio-pragmatic appropriateness (McNamara & Roever, 2007), in other cases it is mainly considered in terms of communicative effectiveness, i.e. success of information transfer (Upshur & Turner, 1995) or successful task completion, i.e. relevance and effectiveness of content according to task instruction (De Jong et al. 2007, 2012; Kuiken, Vedder & Gilabert 2010, Pallotti 2009). There is no unanimity either as to how functional adequacy can best be assessed. Moreover, it is not clear by which textual and linguistic features functional adequacy is mainly determined in the eyes of raters (however see Iwashita et al. 2008).

Next to this general and language-specific discourse competence, functional adequacy requires mastery of vocabulary and syntax. The high correlations which have been found, particularly for advanced learners, between functional adequacy on the one hand and

lexical diversity and accuracy on the other, suggest that the development of functional adequacy and linguistic complexity may often go hand in hand (Alanen et al 2010; Kuiken, Vedder & Gilabert 2010).

The main research question addressed in the study concerns the investigation of functional adequacy as a component of L2 proficiency. More specifically, the study focuses on the following questions:

1. Which measures of functional adequacy can be inferred from the literature? And how robust are these measures when they are put to the test?
2. How does functional adequacy relate to linguistic complexity and how can functional adequacy be assessed?
3. How can differences between low-level and high-level learners be described in terms of functional adequacy?
4. What are the communalities and differences between L2 and L1 writers with respect to functional adequacy?

These questions allow comparisons between: L2 learners with different proficiency levels; and between L2 and L1 learners. In this way the study may contribute to insights into learnability issues like: the acquisitional path L2 learners follow (from level A2 to C2), differences between L2 and L1 learners in the processing of functional adequacy. It also tests specific assumptions in the development of functional adequacy, such as the 'omega-shaped behaviour' (as suggested by Norris & Ortega (2009) and Pallotti (2009)).

References

Alderson, J. C. (2007). The CEFR and the need for more research. The Modern Language Journal, 91(4), 659–663.

Council of Europe. (2001). Common European framework of references for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R. & Hulstijn, J. H. (2007). The effect of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele, A. Housen, M. Pierrard, F. Kuiken, & I. Vedder (Ed.), Complexity, accuracy and fluency in second language use, learning and teaching (pp.53–63). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

De Jong, N.H., Steinel, M.P., Florijn, A., Schoonen, R. & Hulstijn, J.H. (2012). The effect of task complexity on native and non-native speakers' functional adequacy, aspects of fluency, and lexical diversity. In A. Housen, F. Kuiken, & I. Vedder (Ed.), Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA, Amsterdam/Philadelphia: John Benjamins Publishing Company, 121-142.

Housen, A., Kuiken, F. & Vedder, I. (Ed.), (2012). Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA, Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantatative and qualitative dimensions of language proficiency. The Modern Language Journal, 91, 663–667.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? Applied Linguistics, 29(1), 24–49.

Kuiken, F., Vedder, I. & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin & I. Vedder (Ed.), Communicative

proficiency and linguistic development: Intersections between SLA and language testing research, Eurosla Monographs 1: 81-100.

Little, D. (2007). The Common European Framework of reference for language perspectives on the making if supranational language education policy. The Modern European Language Journal, 91, 644–652.

McNamara, T. F. & Roever, C. (2007). Testing: The social dimension. Malden, MA/Oxford UK: Blackwell Publishing.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. Special issue, complexity, accuracy and fluency (CAF) in second language acquisition. Applied Linguistics, 30(4), 590–601.

Upshur, J. A. & Turner, C. E. (1995). Constructing rating scales for second language tests. ELT Journal, 49(1), 3–12.

Wolfe-Quintero, K., Inagaki, S. & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy and complexity. Honolulu, Hawai'i: University of Hawai'i Press.

Kasper Maes

Radboud in'to Languages, Nijmegen, The Netherlands

k.maes@let.ru.nl

CEFR Grammar: Which Rules at Which Level?

Bio data

Kasper Maes studied German Language & Culture and Philosophy at Radboud University Nijmegen, Netherlands. In the past ten years, he has worked as a German and Dutch language trainer at two different university language centers (Radboud in'to Languages, Tilburg University Language Center). As a project coordinator, he has also been involved in (mostly German) school book projects and online learning websites. Kasper Maes is the author of the German CEFR grammar that was published in 2011 in the Netherlands by Van Dale Publishers.

Abstract

In 2011, a series of grammar books were published in the Netherlands (Van Dale Uitgevers, 2011). These publications were the result of a project, in which the grammar of six different languages (Dutch, English, German, French, Italian and Spanish) was classified according to the CEFR. This project was a response to the need for concretization of the CEFR level descriptors. First, existing publications regarding the CEFR and grammar were consulted, such as the Profile Deutsch for German and the British Council EQUALS study and Pearson PET General Handbook for Teachers for English. Secondly, the CEFR level descriptors were interpreted for indications of the kind of grammar that should be familiar at a particular CEFR level. A third guideline in the project was the complexity and frequency of particular grammatical structures in sample texts. Based on these three guidelines and the experience of the authors, the grammar rules were classified according to the CEFR. In addition, the language in which the rule itself was explained (easier wording, combining at higher levels) and the difficulty and content of the examples (more difficult vocabulary, longer sentences, different language contexts) were also adapted to the CEFR levels. In this presentation, an overview of the project and its results will be given and some of the choices that were made and the difficulties that were faced will be discussed.

Short paper

In the summer of 2010, a group of six authors attended a meeting organized by the Dutch publishing company Van Dale. They discussed the idea to write a set of grammar books for six different languages: Dutch, English, French, German, Italian and Spanish. In these grammar books for the first time the grammar rules would be presented at different CEFR levels, thus making it possible for language learners to look up the rules at their own language level.

In the beginning of the project some choices were made. The primary target group of the CEFR grammar books were adult language learners. However the grammar would also have to be used as a reference work in secondary education. Furthermore, it was decided that the grammar could not be a contrastive grammar. The reason for this was that the option had to be left open to publish the grammar in other countries. Finally, the grammar books had to be reference works in which grammar rules could be looked up, not grammar course books or methods.

The real challenge of this project was to divide the grammar rules using the CEFR levels. In the level descriptors, the can do statements, there is no explicit information on grammar or grammar rules. On the other hand, language learners have to understand certain grammatical rules and constructions to be able to express themselves at a certain CEFR level. Immediately two questions arose: must a learner have implicit or explicit knowledge of the grammatical rule? And does this knowledge equally apply to all skills (reading, writing, listening, speaking)? In order not to make the project an almost impossible task, the authors decided to let this last question be and to focus on explicit grammatical knowledge.

Would all the six CEFR levels have to be taken into account? This question was answered differently by different authors. Most of them agreed on four CEFR levels, A1, A2, B1 and B2. The reason for this was that at the C1 and C2 level, language learners (almost) only acquire specific, highly complex vocabulary structures and sentences. Grammar knowledge does not play an important role anymore. Or in other words, at B2 level, language learners already have to possess all the grammatical knowledge in order to be able to express themselves at the higher C1 and C2 level.

The authors used four guidelines to divide the grammar rules according to the CEFR. Firstly, publications in which grammar was linked to the CEFR levels were used, e.g. Profile Deutsch (Glabionat et al., 2005), Advies-grammaticaleerlijnen Duits en Frans (recommended grammar learning trajectory German and French) (Meijer, 2006) and British Council - AEQUALS: A Core Inventory for General English (North et al., 2010). These publications were taken as a starting point for the project.

Secondly, the authors turned to the can do statements as described in the CEFR (Council of Europe, 2001) and in other publications (Liemberg & Meijer, 2004; Meijer, 2007). To some extent, these statements can be translated into grammatical structures and rules. An example is the B2 statement 'I can speculate about causes, consequences and hypothetical situations'. In English, this statement refers to a sentence like 'Their train might have been delayed'. This leads to the conclusion that explicit knowledge of the modal verb 'might' is presupposed at B2 level. In German, this same statement refers to the use of Konjunktiv II. This subjunctive mood is used to express hypothetical situations.

A third guideline in the project was complexity and frequency. This guideline was used to link complex passive sentence constructions to the B2 level or to distinct between more or less frequent modal verbs at different levels. A clear example for German is the use of the four cases. Frequent cases like Nominativ and Akkusativ are introduced at A1 level. Lesser frequent or more complex cases like Dativ and Genitiv are explained at A2 and B1 level.

Sometimes these three guidelines did not lead to a satisfactory answer to the question: to which level does this rule apply? In this case the authors used their experience as a fourth guideline to make specific choices.

On the basis of these four guidelines the grammar books (Van Dale Uitgevers, 2011) were written. Not only were the grammar rules linked to the CEFR levels, but the complexity of both the rules and language examples was adjusted to the CEFR levels. In other words, a basic A1 grammar rule was extended at A2 level and illustrated by a more complex A2 sentence.

The grammar project was finished in the beginning of 2011. The books were published in the autumn of that same year.

This CEFR grammar project is linked to the conference topic 'Competence and performance', because it tries to make the can do statements more explicit in terms of

grammar knowledge. Although the CEFR can do statements do not provide any explicit information about grammar or grammar knowledge, the four guidelines mentioned above make it possible to link grammar rules to the different CEFR levels. This project is one of the first attempts to provide a detailed grammatical framework which can be used by language learners who are familiar with the CEFR.

References

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_en.pdf.

Glabionat, M., Müller, M., Schmitz, H., Rusch, P. & Wertenschlag, L. (2005). Profile Deutsch, Berlin usw: Langenscheidt.

Liemberg, E. & Meijer, D. (2004). Taalprofielen. Retrieved from http://www.slo.nl/downloads/archief/Handreiking_20MVT_20Duits_2c_20Engels_2c_20Frans.pdf/

Meijer, D. (Red.). (2006). Advies-grammaticaleerlijnen Duits en Frans. Enschede: SLO. Geraadpleegd via http://www.slo.nl/downloads/archief/Advies.pdf

Meijer, D. & Fasoglio, D. (2007). Handreiking schoolexamen moderne vreemde talen havo/vwo. Enschede: SLO.

North, B., Ortega, A. & Sheehan, S. (2010). British Council - AEQUALS: A Core Inventory for General English. Retrieved from http://www.eaquals.org/cefr/

Van Dale Uitgevers. (2011). Van Dale Grammatica Engels / Duits / Frans / Spaans / Italiaans / Nederlands (NT2): Glashelder overzicht op elk taalniveau. Utrecht.

Margret Oberhofer & Jozef Colpaert

Universiteit Antwerpen, Antwerpen, Belgium

margret.oberhofer@ua.ac.be – jozef.colpaert@ua.ac.be

Language for Specific Purposes and the CEFR – the EuroCatering.org Example

Bio data

Margret Oberhofer is the International Projects Co-ordinator within LINGUAPOLIS, the Institute for Language and Communication at the University of Antwerp. Since 2005, she has been and still is coordinating and implementing several national and international R&D projects related to e-learning and language learning. For an overview of the different projects please go to: http://europeanprojects.linguapolis.be. Margret is also a part-time eLearning assistant at the Institute for Education and Information Sciences (IOIW) at the University of Antwerp.

Jozef Colpaert is Professor of Educational Engineering, Instructional Technology and Computer Assisted Language Learning (CALL) at the University of Antwerp. He develops courseware applications and has carried out research projects in language pedagogy and courseware design. Furthermore, he is the director of Research and Development at LINGUAPOLIS. He is also the editor-in-chief of the CALL Journal (Taylor & Francis) and organiser of international CALL conferences.

Abstract

www.eurocatering.org is a web-based language learning tool in 12 languages designed for trainees, students and workers in the Hotel and Catering industry to improve their oral language skills. EuroCatering helps the envisaged target groups acquire the basic specific vocabulary and the communicative competences needed to function efficiently in a kitchen, restaurant or hotel abroad by providing learning materials and instructional support. The first part of the presentation discusses the challenges related to a language course for specific purposes to a certain level of the Common European Framework of Reference for Languages (CEFR), however, the CEFR does not refer to languages for specific purposes (Alderson, 2007; Krumm, 2007; Komorowska, 2012). We will also discuss the Reference Level Descriptions (RLDs), which is a new generation of reference descriptions currently available in ten languages.

The second part of the presentation highlights the EuroCatering Language Portfolio as a way to provide a self-assessment tool for students of this vocationally-oriented language learning course.

The presentation closes with consideration of the practicality of the CEFR and the RLDs for developers of an online language course for specific purposes.

Short paper

www.eurocatering.org is a web-based language learning tool in 12 languages designed for trainees, students and workers in the Hotel and Catering industry to improve their oral language competences. EuroCatering helps the envisaged target groups acquire the basic specific vocabulary and the communicative competences needed to function efficiently in a kitchen, restaurant or hotel abroad by providing learning materials and instructional support.

The first part of this paper describes the challenges related to a language course for specific purposes to a certain level of the Common European Framework of Reference for Languages (CEFR) and the Reference Level Descriptions (RLDs). The second part of the paper presents the EuroCatering Language Portfolio as a way to provide a self-assessment tool for students of this vocationally-oriented language learning course. The paper closes with consideration of the practicality of the CEFR and the RLDs for developers of an online language course for specific purposes.

Introduction

www.eurocatering.org is a web-based language learning tool in 12 languages1 designed for trainees, students and workers in the Hotel and Catering industry to improve their oral language competences. EuroCatering helps the envisaged target groups acquire the basic specific vocabulary and the communicative competences needed to function efficiently in a kitchen, restaurant or hotel abroad. In the context of EuroCatering, the trainees are 'social agents' that need to function adequately in a stressful working environment and understand orders and tasks given by the manager or the chef. It focuses mainly on receptive skills and very basic productive language skills, such as following short instructions and answering short questions. The aim of the course is to develop a specific, linguistic repertory or a partial competence rather than achieve mastery in certain languages. EuroCatering Language Training is the result of a European project; it currently has 9000 committed learners from all over Europe and beyond.

In a needs analysis conducted at the beginning of the project it became apparent that the first work placement abroad is a crucial period in the life of the trainees, as it may determine their view of pursuing a career in the catering sector. Basic language skills can help make this experience successful. Results from 35 in-depth interviews suggest that the trainees mainly want to "do" practical things rather than learn theory such as grammar rules or read lengthy texts. A language tool helping them to acquire basic language skills had to be visual and practice-oriented. Based on this, the three design objectives for EuroCatering were to develop a language course with a) relevant and specific vocabulary for a kitchen and restaurant environment that avoided theory and too much emphasis on written text, b) that is visually attractive and engaging as this specific target group is not particularly interested in studying a language but prefers a practical approach and c) that is easily accessible for all language learners, including those with no knowledge of the target language.

The course consists of 34 sections divided into topics such as "at the pass" and "complaints and compliments" containing 411 exercises and 68 interactive dialogues. The development of these materials took place over three phases: at the beginning, the exercises, dialogues and vocabulary were designed in English by language teachers familiar with the subject. All exercises and phrases were revised and simplified to avoid complex structures (e.g. passive voice, subordinate clauses). During the second phase, the pilot material was translated into French. Again, exercises and phrases with complex structures were simplified in English if translation proved difficult. At the end, the final English and French versions were used as the basis for translating into the other ten languages.

When developing the course, the team had the basic user (A1- A2) of the Common European Framework for Languages (CEFR) in mind. However, it proved to be challenging to design a Language course for Specific Purposes (LSP) that fits completely into the basic level of the CEFR and provided sufficient and relevant terminology for learners of the catering sector. According to Alderson (2007), the reason for this is "that in its current form the CEFR is not suitable for young learners, for the teaching of languages for specific purposes, or for CLIL" (p.662). Krumm (2007) echoes this by

[1] The 12 EuroCatering-languages are: Dutch, English, Finnish, French, Galician, German, Irish, Italian, Norwegian, Polish, Slovenian, and Spanish.

pointing out that the vocational and administrative environments are not currently the focus of the CEFR. Komorowska (2012) referred to the lack of clear and sufficient references to LSP, especially when evaluating proficiency, with "no more than a few examples from the Threshold Level (CEFR: 26-27)" and "some references in the self-assessment grid (CEFR: 26-27), though only at B2-C1-C2 levels" (p.110).

The Common European Framework of Reference for Languages (CEFR)

The Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR) was launched in 2001 by the Council of Europe . It was designed to provide a transparent, coherent and comprehensive basis for the elaboration of language syllabuses and curriculum guidelines, the design of teaching and learning materials, and the assessment of foreign language proficiency. Representing a supranational organisation, the Council's goal was to overcome the barriers to communication amongst professionals in the field and promote international co-operation "by providing them with a common basis for the explicit description of objectives, content and methods" and thus "enhance transparency of courses, syllabuses and qualifications" (Council, 2001, p.1). The CEFR is used in Europe but also in other continents and is available in 38 languages.

The CEFR comprises "(a) a descriptive scheme for analysing what is involved in language use and language learning and (b) a definition of communicative proficiency at six levels arranged in three bands — A1 and A2 (Basic User), B1 and B2 (Independent User), C1 and C2 (Proficient User)" (Little, p.645). The CEFR is a non-language specific framework which uses the same descriptors, for example, for Polish, Greek and English. Its descriptive scheme embraces general and communicative language competences. The general competences of language learners consists of their declarative knowledge (savoir), skills and know-how[2] (savoir-faire), existential competence (savoir-être) and the ability to learn (savoir-apprendre). The communicative language competences comprise the linguistic, sociolinguistic and pragmatic components.

The CEFR distinguishes four domains: the public domain, the personal domain, the educational domain and the occupational domain. The latter is the most relevant for EuroCatering, embracing "everything concerned with a person's activities and relations in the exercise of his or her occupation" (Council, 2001, p.15). The CEFR proposes the notion of partial competences, such as learning a foreign language in order to perform to a higher standard at work.

The approach adopted by the CEFR is action-oriented, meaning a language is learnt for a social purpose (North, p.656). The learners are therefore social agents, "i.e. members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action" (Council, 2001, p.9).

The Reference Level Descriptions (RLDs)

In order to meet the needs of teachers, textbook authors and operators who considered the CEFR as too broad, launched a new generation of reference descriptions. While the CEFR was drafted without reference to any specific language, the new generation of reference descriptions are drafted for specific languages and are commonly known as Reference Level Descriptions (RLDs) for national and regional languages.

Currently, a list of RLDs is being developed or has already been finalised for the following ten languages: Czech, German, English, French, Georgian, Greek, Italian, Norwegian, Portuguese and Spanish. Unfortunately, these languages only constitute half of the EuroCatering languages.

[2] They include: Vocational and professional skills: the ability to perform specialised actions (mental and physical) required to carry out the duties of (self-) employment. (Council, p.104)

The RLDs for English are collected within the English Profile Programme, a long-term research programme sponsored by the Council of Europe, and based on research using the Cambridge Learner Corpus. The English Vocabulary Profile (EVP) is part of the English Profile Programme and its database can be consulted online, where it is readily available for the time being. The EVP shows which words and phrases learners around the world know at each level of the CEFR in both British and American English. The consortium behind the EVP emphasises that rather than providing a syllabus of the vocabulary that learners should know, the EVP project verifies what they do know at each level. The data collection includes English students worldwide and is currently on-going.

When applying the EVP to a set of vocabulary of EuroCatering.org it became apparent that it is not possible to develop a language course for specific purposes within a certain CEFR level only. Table 1 illustrates a) that the vocabulary is spread over all six levels, with an emphasis on A1-A2, b) that some words have different levels (see variable) and c) that some words have not yet been labelled.

A1	kitchen, doctor, juice, knife, eleven
A2	dish, dessert, jam (sweet food), kilo, garlic, hundred
B1	delicious, dust, diet (usual food & weight loss)
B2	dairy (products)
C1	compliment (verb) C1
C2	cutlery
Variable[3]	keep (to have): A2; keep (food): B2 do (verb, action): A1; do (verb, make): A2; do (auxiliary verb, avoid repeating): B1; do (auxiliary verb for emphasis): B2 decide (verb, choose): A2; decide (verb, result): C2 repeat (verb, say again) : A2; repeat (verb, tell): A2 slice (verb): B2; slice (noun, food) A2 compliment (verb) C1; compliment (noun, praise): C2
Not found	dressing, demi-glace, deep fryer, simmer

Table 1: CEFR Levels of the EuroCatering vocabulary according to the EVP

The EuroCatering Language Portfolio

EuroCatering consists of two main parts: The Cloche and The Tray. The Cloche presents the language course in a safe and familiar environment, while The Tray provides supporting tools such as a professional dictionary, intercultural information, a teacher's corner and the EuroCatering Language Portfolio. It is a self-assessment checklist designed by the team with descriptors developed specifically for EuroCatering.

Attempts to define descriptors of language competences for the technical language in the catering and tourism sectors already began in a Leonardo da Vinci mobility project in 2002/3. The idea of a portfolio for technical language was considered a good idea, but for several reasons it proved too challenging to fully develop within such a short timeframe. Within the EuroCatering project (Leonardo Da Vinci, 2006-2008), the development of the European Language Portfolio (ELP) tailored to the specific needs of the target groups (students and teachers in vocational education, and workers in catering SMEs) was one of the promised results.

The task of the EuroCatering Language Portfolio is to monitor, guide and motivate students during their learning process. Self-assessment is not only a tool for motivation but also to raise awareness by "helping learners to appreciate their strengths, recognise their weaknesses and orient their learning more effectively" (Council, 2001, p.192).

[3] The EVP assigns CEFR levels not just to the words themselves, but to each individual meaning of these words. So, for instance, the word *degree* is assigned level A2 for the sense TEMPERATURE, B1 for QUALIFICATION, B2 for AMOUNT and C2 for the phrase *a/some degree of (sth)*.

Against scepticism towards self-assessment, Little (2007) opposes that learners "quickly develop the ability to assess their own learning [with self-assessment-tools], at least in terms of what they can do in the target language(s). One obvious way of ensuring a minimum level of validity is to require them always to prove that they can do what they claim" (p. 651).

Ideally, the self-assessment tool can function as a complement to tests and teacher assessments conducted in vocational schools. This might be possible for foreign languages such as English, French, Spanish, German and possibly Italian. However, the majority of foreign languages on EuroCatering.org such as Norwegian, Slovenian, Galician, Polish, Finnish, Dutch or Irish are not taught in vocational schools. The combination of languages that are not frequently taught in schools around Europe and the specific terminology needed in these languages make it very unlikely that there are assessment tools available for these specific cases. In this particular situation the EuroCatering Language Portfolio can provide a form of assessment.
Like the ELP, the EuroCatering Language Portfolio consists of three parts:

- The Language Passport, where students can fill out their language skills for three topics: basic vocabulary, specific vocabulary used in the kitchen and vocabulary used in the restaurant. The idea is that students check from a list of descriptors stating if they understand or speak the language.
- The Language Biography, a) with personal information, b) a list to include relevant documents such as school certificates, attestations or the Europass, c) a short-term learning plan where students can define their learning objectives and if they have achieved them or not, and d) the self-assessment checklist.
- The Dossier that can be filed by the students with a list of documents, websites and a European CV.

The self-assessment plan is subdivided into 4 parts:

1. Basic vocabulary with eight topics, such as numbers and uniforms. Students are required to tick the box if they can understand (U) and/or say (S) something specific.

Topic	**Descriptor**	**U**	**S**
Numbers	I can understand/say numbers from 1-100.		
	I can understand/say numbers from 100-1000.		
Uniforms	I can understand/say words describing uniforms for kitchen and restaurant staff, such as *jacket, pants, apron, hairnet.*		

Table 2: Examples of the self-assessment checklist for basic vocabulary

2. The kitchen vocabulary includes descriptors related to 14 topics, including fish or vegetables and 12 communication settings, such as cold room and fruit preparation.

Topic	**Descriptor**	**U**	**S**
Fish	I can understand/say words indicating sea fish, such as *hake, tuna, sole.*		
Vegetable	I can understand/say words indicating vegetables, such as *cabbage, leek, carrot.*		
Communication			
Cold room	I can understand/say simple commands related to keeping food and temperature records in the cold room.		
Fruit preparation	I can understand/say simple commands related to most frequently used methods of fruit preparation, for example, *remove the seeds, cut the grapes in half.*		

Table 3: Examples of the self-assessment checklist for kitchen vocabulary

3. The restaurant includes eight topics, including menus and billing and 16 communication topics such as cutlery and greeting and seating.

Topic	**Descriptor**	**U**	**S**
Menus	I can understand/say words indicating different types of menus, *such as children's menu, menu of the day.*		
	I can understand/say words related to the structure of the menu, for example, *cold first course, hot first course, main dish.*		
Billing	I can understand/say words and expressions describing different methods of payment, such as *cash, credit card, bank card.*		
	I can understand/say expressions frequently used in billing, such as *pay, mistake, too much, enter your pin.*		
Communication			
Cutlery	I can understand/say commands about placing and using cutlery.		
Greeting and seating	I can understand/ask questions related to meeting and seating the guests, for example, *Have you booked? May I take your coat? Would you follow me, please?*		
	I can understand/respond to the guests' questions regarding seating/table availability.		

Table 4: Examples of the self-assessment checklist for restaurant vocabulary

4. Intercultural information with ten topics, including greetings, culture at work etc. This section is interesting for trainees doing an internship abroad.

Topic	**Descriptor**	**U**	**S**
Greetings	I can understand/say and respond to formal and informal greetings.		
Culture at work – restaurant	I know the hierarchical structure of the staff.		
	I know specific ways of serving and removing dishes.		
	I know/understand and can say the main regulations connected with work in the restaurant, for example, Hazard Analysis Critical Control Points (HACCP).		

Table 5: Examples of the self-assessment checklist for intercultural information

These are descriptors of language proficiency and language competence that cannot be found in the CEFR as they are specific for the EuroCatering context. As these descriptors are so specific to this sector, it was more important that the EuroCatering Language Portfolio "develops the learners' self-assessment skills than to establish reliable links between their language proficiency and the common reference levels" (Lenz & Schneider, 2000, p.1).

The descriptors are, as suggested by the Council of Europe a) formulated in a positive rather than a negative way, b) describing concrete tasks and/or concrete degrees of skill in performing tasks, and c) kept short (max. 25 words) (Council, 2001, pp.205-207). The suggestions to avoid jargon and complex syntax in the portfolio is only partly implemented as the language course itself consists of specific terminology.

Conclusion

This paper describes the challenges within the EuroCatering European project to develop an online course for the catering industry that is in line with a specific level of the Common European Framework (CEFR). It describes the content of the EuroCatering Language Portfolio as a self-assessment tool that helps students to monitor and guide

them throughout the learning process for specific language courses. In the concrete case of EuroCatering.org the CEFR and the European Language Portfolio provided a guideline when implementing the course and the self-assessment tool. It enhanced communication related to the objectives and content amongst the project team members from all over Europe. However, in the case of EuroCatering, an online course for language for specific purposes designed for a specific target group, the CEFR was less applicable as it might be in other language learning situations.

References

Alderson, J.C. (2007). The CEFR and the Need for More Research. The Modern Language Journal, 91, 659–663.

Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). Retrieved from http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe. (Update 2011). Reference Level Descriptions (RLD) for national and regional languages. Retrieved from http://www.coe.int/t/dg4/linguistic/dnr_en.asp
Douglas, D. (2000). Assessing Language for Specific Purposes, Cambridge University Press.

English Vocabulary Profile. Retrieved from
http://vocabulary.englishprofile.org/staticfiles/about.html

European Language Portfolio. Dedicated website. Retrieved from:
http://www.coe.int/t/dg4/education/elp/

European Language Portfolio. Retrieved from:
http://www.coe.int/t/dg4/linguistic/Portfolio_en.asp

Komorowska, H. (2012). Academic Perspectives from Poland. In Byram, M.; Parmenter, L. (Eds.), The Common European Framework of Reference. The Globalisation of Language Education Policy, 104-113.

Krumm, H-J. (2007). Profiles Instead of Levels: The CEFR and Its (Ab) Uses in the Context of Migration. The Modern Language Journal, 91, 667-669.

Lenz, P. & Schneider, G. (2000). Introduction to the bank of descriptors for self-assessment in European Language Portfolios. Retrieved from
http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Intro_descriptors_bank_EN.pdf

Leonardo da Vinci – Mobility/Exchange (2002/2003). Technical Language in Catering in Tourism. Summary of project results. Secondary Vocational School of Catering and Tourism Celje, Slovenia.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspective on the Making of Supranational Language Education Policy. The Modern Language Journal, 91, 645 – 655.

North, B. (2007). The CEFR Illustrative Descriptor Scales. The Modern Language Journal, 91, 656– 659.

Reference Level Descriptions (RLDs). Retrieved from http://www.coe.int/t/dg4/linguistic/dnr_en.asp
Štiherl, D. (2005). Fachsprachenportfolio (FSP) für Gastgewerbe und Tourismus. Festlegung der Schlüsselkompetenzen im berufsbezogenen Fremdsprachenunterricht. IDT-Konferenz, Graz.

Harold Ormsby

Universidad Nacional Autónoma de México, Mexico City, Mexico

haroldormsby@yahoo.com

Self-Assessment as a Starting-Point for Useful Communication among Learners, (Prospective) Employers and Teachers: Adapting the European Language Portfolio (ELP) for Use in Specific Real-World Contexts

Bio data

Harold Ormsby has been a language teacher (L1-L2, L2-FL and FL) for 50 years. His language-learning and language-teaching experiences have covered the history of the field from grammar-translation and Audio-Lingual clear through to whatever it is we are doing nowadays. His interest in assessments of all kinds came early because of his consistent inability to do well on tests, effort notwithstanding. His focus has always been on fairness for test-takers and the professionals who help them prepare for tests. His involvement in the training of teachers of Mexican and other indigenous languages strengthened his view that assessment of all kinds in educational and professional settings is rooted in the varied cultural, social and economic mixes each of us lives in. This is especially true for the assessment of language proficiencies because neither in social groups nor in individuals can languages, cultures and social patterns be winnowed apart. His interest in languages in international professional settings started with an invitation to advise pilots about the real-life implications of the then newly imposed aviation language proficiency requirements. This meshed with a long-standing interest in Legal English, and this in turn linked with International Legal English (ILE), in which he has held a certificate since 2010. His current work has to do with teaching ILE to users of Spanish, and, in some cases, helping them prepare for the certificate examination. As part of this, he is exploring ways of fostering useful communication between potential employers and potential employees about language proficiencies.

Abstract

People often say they are learning an additional language for employment purposes. Indeed, many employers want to have or find employees who are proficient in (a) certain language(s), notably English. However, most often language teachers deal with prospective employees who cannot know who their employer(s) will be. Textbook writers, in turn, bridge gaps by taking generic employment contexts as their guides. Furthermore, employers very, very seldom communicate directly and usefully with teachers or (not surprisingly) with prospective employees about language proficiency expectancies, demands, needs etc.

The CEFR and the subsequent ELP have the goal of improving "the quality of communication among Europeans of different language and cultural backgrounds ... because better communication leads to freer mobility and more direct contact, which in turn leads to better understanding and closer co-operation." (CEFR, pp. xi-xii) Improving communication among learners, (prospective) employers and teachers is, we believe, essential to making this universally applicable goal achievable.

The long-range objective of this design-development project is to create a multidimensional assessment scaffolding (array, matrix, framework) in/on which specific-

purpose can-do statements can be located, using categories, classifications and criteria that are found (perceived) and make sense in the real world, and that can be modified as the real world changes.

The project's current reality is preparing Mexican adults with English as L1-L2 or FL to take the International Legal English Certificate (ILEC) examination and/or who have professional needs and interests related to International Legal English. Examples to explain the underlying idea and to show how the current scaffolding works in practice will come from self-assessments used before, during and after exam preparation, and in the creation and updating of individuals' ELP-like language autobiography and language-experience dossier. One hopes to find other language teachers who would like to do similar projects in their own languages and realities.

Short paper

Whence and whither

What we (professor Diana Jenkins and I) are doing with International Legal English (ILE) is entirely exploratory. Explorations are not aimless wandering. Like any explorers, we think we know where we are going but we are not sure if, how or, even less, when we might get there and we do not know what adventures await us along the way. We expect to make mistakes and we hope to learn from them quickly and clearly. We fully expect that our destinations will, in fact, be dynamically moving goalposts; the spirit of the times, needs and available resources change, and each change moves a goalpost even if it is only vaguely seen in the distance at best.

The adventure began quite a few years ago when one or another of us was assigned to give a business or legal English course as one of the many one-semester, unsequenced advanced courses our university's foreign language center regularly offers. Course by course, we built up a relatively disorganized but also rather detailed panorama of students' present realities, immediate needs, possible needs, wishes and desires, strengths and weakness from a language-learning point of view, and so on. Eventually, we prepared ourselves to take (and ultimately do well on in 2010) the examination for the International Legal English Certificate (ILEC). We had been sorely tempted to advise our students to work towards the Certificate but we felt we should not do so until we had experienced the examination "from the inside." At the same time, we developed a vision, course-, lesson- and textbook-use- plans, and materials (most for online delivery) for preparing our students both for the very limited goal of receiving the certificate and the broader and, in the end, much more meaningful goal of being successful (plurilinguals-pluriculturals) in their profession and its labor market.

Neither we nor our students are Europeans. From a strictly scientific point of view, this point probably is not very important; however, from the point of view of down-to-earth practicalities and realities, it is very significant. Both our students who have jobs now in Mexico and those who are or soon will be seeking jobs anywhere they can find them, can feasibly need to make themselves professionally attractive to a daunting variety of potential employers. Deliberately overemphasizing or ignoring potential employers from any given country, region, industry, legal specialty and so forth, would be a serious ethical mistake on our part. Making believe that we can usefully stereotypify all these employers is equally foolish, as it would be if we were to dream of having direct, face-to-face contact with any but a very, very few potential employers.

Our decision to use can-do statements (C-DS), gathered into a dynamically changing (hopefully, advancing) setting-specific (ILE) corpus as the underlying building materials of initial and periodic individual goal-setting, courses, lessons, exercises and on-going assessments will not be discussed here. C-DS are part of a long tradition in Applied Linguistics. Although, clearly, there are theoretical (and even some practical) topics that are under discussion in the field, we have decided to set those aside for the foreseeable

future. Likewise, at this point, we do not feel that a major discussion of the down-to-earth practicalities of the ELP is needed in our context. Europe has made and is making proposals for their own ends. What we have done is to propose to our students using any unpreformatted design, inspired in any European or non-European models, to bring together presentable language autobiographies and dossiers ("collections-of-stuff") that may well be informative for potential employers.

We have also decided not to do any formal research into how big the mess of communication among language-learners, language-teachers and potential-employers (as types) is, or what the mess really looks like "from the outside" or "from above." Assuming it is a mess does not preclude discovering that at certain points or in some areas things are not in fact all that messy; and of course, we have no intention of deliberately making messes any messier or making a mess where there is none. Furthermore, we rather think that whatever messes there are are largely products of the facts:

1. that employers talk about L1, L2 or FL use, L1, L2 or FL learning, and/or L1, L2 or FL teaching in terms of their own personal experiences in school (at what ages?) and in life;
2. that students see employers as dangerous, unpredictable demi-gods, and at times apply the same attributes to teachers;
3. that students also talk about L1, L2 or FL use, L1, L2 or FL learning, and/or L1, L2 or FL teaching in terms of their own personal experiences in school (at what ages?) and in life;
4. that teachers themselves also talk about L1, L2 or FL use, L1, L2 or FL learning, and/or L1, L2 or FL teaching in terms of their own personal experiences in school (at what ages?) and in life but add myriad influences from their experiences and training as teachers, none of which can be shared (as peers) with either of the other two types of actors; and finally,
5. that everyone concerned (including ourselves) cannot escape the effects and influences of our societies' beliefs, attitudes and practices in relation to languages;

all of which constitute an ideal recipe for totally ineffective interpersonal and intergroup communication.

Fostering communication will not be possible until there is a more or less generally accepted way of getting people to talk about language use and proficiencies in ways that will help, eventually, each of the actor-types (employers, students, teachers) say meaningful, useful things to the other actor-types (stakeholders). This is not going to happen if we in Applied Linguistics only talk it over among ourselves and foist our decisions on everyone else. No one is an expert here, although everyone has a kind of perfectly respectable expertise.

For the reasons given in the next section, we have chosen to use a modification Q-Method (also known as Q-Sort) as an early vehicle for our explorations in the terra incognita of inter-stakeholder communication.

Our short- and medium-term research procedure, and a proposal for those who are reworking the CEFR-ELP

Q-Method has been known in Psychology's bibliography since the mid-1930s, has never become mainstream but is being used more and more frequently for research in a wide variety of fields. It is a procedure for studying subjectivities. For quick-and-easy introductions, see the videos by Deignan (2012) and Glasgow Caledonian University (n.d.). A great deal more information can be had from the Q Methodology, The Q Method and the Applied Qualitative Methods Network webpages, from Brown (1996 and 1993), Donner (2001), Shinebourne (2009) as well as from other more detailed and technical

sources that are not listed in this short paper's reference list but that do appear among the references given by the aforementioned.

During this paper, a hands-on example of doing a (small, quick) Q-Sort, as well as further examples drawn from our current ILE corpus of some 100 C-DS will be provided. This short paper will give an overview of our reasons for choosing and temporarily adapting Q-Method, in order to leave room to briefly outline a proposal for those who want to rework the CEFR-ELP.

There are three aspects of the Q Method that caught and continue to hold our attention: (1) No attempt is made to say that a given C-DS has a meaning that can be determined by the researcher; the research looks for patterns (including blank spaces) in subjects' individual, subjective understanding of C-DS. (2) Subjects do not have to be themselves; they can assume roles (one at a time) and organize the C-DS according to their (subjective) understanding of what, e.g., the person who normally plays that role would do as they sort the C-DS. (3) In an informal or barely formal research session (of the kind we envision), doing a Q-sort can be fun and stimulate a brief, focused, conclusion-oriented conversation about individual C-DS or small (perhaps accidental) groupings of them.

Aspect (1) makes it unlikely that a researcher's own professional experience(s), life experiences, beliefs, prejudices (that is, subjectivity) and/or those subjectivities that come from published teaching programs and materials, test specification tables, socio-political policy dicta, and the like will invade data collection to any prejudicially significant degree. Through their responses, subjects can (and fairly often do) tell a researcher that s/he has "gotten it all wrong" and, at the same time, they can show him/her what "getting it right" may well look like.

Aspect (2) means that, for instance, a single student can do one Q-Sort while playing the role of MYSELF TODAY, another Q-Sort as MYSELF AFTER SIX MONTHS FOLLOWING MY STUDY PLAN, and a third Q-Sort about MY BOSSES' EXPECTANCIES FOR MY STUDY PLAN. Likewise, if one can get a boss's cooperation (not impossible), the boss him/herself can sort the same set of C-DS while playing the roles of, first, MY [the boss's] EXPECTANCIES FOR THIS EMPLOYEE'S STUDY PLAN, second, THIS EMPLOYEE TODAY, and finally perhaps, THIS EMPLOYEE IN SIX MONTHS. Role playing can help teachers look at themselves, at their students and at the world-of-work in new, hopefully more perceptive, ways.

Aspect (3) means that it is relatively likely that research will be done and that small, informal reports will be tremendously informative. Typically, a subject does each Q-Sort on his/her own, while the researcher watches. However, there's no reason why a sort cannot be done by one person who is off-and-on bothered by a couple of others, with the sole purpose of orienting potential subjects to the procedure and listening to their banter about the corpus of C-DS. This banter can be as informative as a formal research session, at least for the next decade or so of development of the language teaching profession.

For the time being, we have adapted certain aspects of the standard Q procedures in order to, first, make it highly attractive to teachers who are not, as a rule, given research time as part of their jobs and, second, to make it more likely that "busy executives" (their term for themselves) will be willing research participants. For teachers, we have eliminated all mention of statistics; at this exploratory stage a great deal of time could be spent on learning to do calculations (with or without a computer) when what the researcher-teacher needs is hand-on experience pulling together and preparing sets of C-DS, getting subjects, writing instructions for them, handling Q-Sort sessions and, importantly, seeing what results look like as text. When they want to, researchers can

start using statistics (including Q-specific programming) but there is a lot to do without them.

For "busy" people, we have eliminated the details of the central (neither-this-nor-that) sector of a Q-Sort setup. Thus, working for instance with a set of 30 C-DS (up to maybe 50), a subject is asked to:

1. Look through all the statements, making quick decisions as to whether any given C-DS should go on the right-hand side (which might mean "bottom-most priority" or "not much priority"), on the left-hand side (perhaps "topmost priority" or "high priority") or in the middle (undefined in our adaptation).

2. Go back to the right-hand pile to decide which 2 (two and only two) statements would go at the very bottom, and which 3 (three and only three) would go one tiny step to the left.

3. Go back to the left-hand pile to do the same: select the 2 statements that should be on the far left-hand edge and the 3 statements that should go just slightly to the right of those.

Subjects are permitted to change their minds as many times as they like. They are invited to look through the middle pile, although without prompting they seem to do that to reconsider decisions they remember making.

In the end, this means that the subject has been asked to make clear decisions about a total of ten (5 to the right, 5 to the left) C-DS. At the same time, the statements that are left in the middle are no less informative. It is normal Q-procedure to interview subjects about their decisions after they have finished.

We believe that as any one loose social group (e.g. students at an institution, coworkers, executives from one corporation) gets used to doing Q-Sorts, the number of C-DS considered can be increased and subjects can be asked to go through the statements in the middle pile in order to push five to the left-of-middle, and another five to the right-of-middle. This would then begin to approximate classical Q-Sort procedure and open the way to statistical analyses but without forcing anyone into them.

We have come up with a way of presenting statements so that each subject's responses can be recorded by taking a picture of their sort with a mobile telephone's camera. The idea is that the photos can be used after a subject as left to prepare a more or less formal record of the subject's final decisions so they can be analyzed. The success (if any) of this experiment will be reported during the paper.

Reports on Q-Sort experiences should be concise narratives that explain what was learned and why/how it was learned and what it means for the real world. This is a topic that will have to be left for another time and place.

To end, some respectful advice to those who are proposing reworking the CEFR-ELP:

First, we strongly recommend looking to Q-Method as a research procedure. If, in the end, you decide that it will not do anything you need, then drop it. But do not do that until you have given it a fair chance.

Second, learn patience. The CEFR-ELP proposals deserve respect even though they undoubtedly need changing.

Third, do not ask for funding. Very, very seldom do funding agencies encourage patience. They tend to want results that are foreseen in a calendar. We believe that one of the

great attractions of Q-Method (and what makes it fully possible for us to be experimenting with it) is the fact that tiny research cohorts and informal, often impressionistic analyses are not only possible but, as often as not, encouraged. Q is cheap.

Fourth, if you have some (or, worse, a lot of) experience with R (i.e. non-Q) methods, it is best to leave that experience to one side for a while as you do hands-on experimenting with Q-Method. There is nothing particularly difficult about Q but it is quite different from R. Research designed for R and research designed for Q cannot be cross-pollinated validly.

Fifth, engage private sector employers and other socially empowered groups. Apparently Europe has already begun this process. Europeans should join in; those of us from "non-Europe" will be wanting to see how you do it. Q may give you a way to pique their curiosity without invoking their biases.

Sixth, engage students but in a way that will keep your social power as teachers from intimidating them. Q can help with this.

And seventh, engage each other. Because Q-Method asks subjects (and, therefore, researchers, as well) to make concrete decisions and encourages them to carry on decision-oriented conversations, intra-group and interpersonal communications can be enlightening and visibly productive. Q discourages unproductive kvetching.

References

Short, introductory videos:

Centre for Canadian Language Benchmarks. (2012). Understanding the ESL Literacy Benchmarks: A concise overview of the Canadian Language Benchmarks and ESL Literacy Phases. (4'29") http://youtu.be/X01IjmN9VM4 (Last accessed 19 March 2013).

Deignan, T. (2012). Leeds Metropolitan Quick-Q Animation (5'22"). http://youtu.be/0AejeH6jw2c (Last accessed 19 March 2013).

Glassgow Caledonian University. (n.d.) Rachel Baker Introduces Q Methodology. (7'51") http://youtu.be/ZbZ2Kq-Fzxo (Last accessed 19 March 2013).

Broadly informative webpages:

Applied Qualitative Methods Network (AQMeN). http://aqmen.ac.uk/ (Last accessed 19 March 2013).

Centre for Canadian Language Benchmarks/Centre des niveaux de compétence linguistique canadiens. http://www.language.ca/ (Last accessed 19 March 2013).

European Centre for Modern Languages (ECML/CELV). European Language Portfolio: Whole School Use (ELP/WSU). http://elp-wsu.ecml.at/ (Last accessed 19 March 2013).

EuroPortfolio. [Online, open source initiative; proposal for the European Commission] http://www.europortfolio.org/ and http://www.eportfolio.eu/ (Last accessed 19 March 2013).

International Legal English Certificate (ILEC). http://www.cambridgeenglish.org/exams-and-qualifications/legal/ (Last accessed 19 March 2013).

Q Methodology: A method for modern research. http://qmethod.org & International Society for the Scientific Study of Subjectivity (ISSSS). http://qmethod.org/issss & Discussion group Q-METHOD. (LISTSERV@LISTSERV.KENT.EDU) (Last accessed 19 March 2013).

Schmolck, P. The Q Method Page. http://schmolck.userweb.mwn.de/qmethod/ (Last accessed 19 March 2013).

Test of Legal English Skills (TOLES). http://www.toleslegal.com/ (Last accessed 19 March 2013).

Articles and books:

Basturkmen, H. (2010). Developing Courses in English for Specific Purposes. London: Palgrave Macmillan, London.

Brown, S. R. (1980). Political Subjectivity: Applications of Q Methodology in Political Science. New Haven & London: Yale University Press. Open access: http://qmethod.org/papers/Brown-1980-PoliticalSubjectivity.pdf

Brown, S. R. (1993). A primer on Q methodology. Operant Subjectivity, 16(3/4), 91-138. Author's version available at http://facstaff.uww.edu/cottlec/QArchive/Primer1.html

Brown, S. R. (1996). Q methodology and qualitative research. Qualitative Health Research, 6(4), 561-567.
Available at Schmolck http://schmolck.userweb.mwn.de/qmethod/srbqhc.htm

Centre for Canadian Language Benchmarks. (October 2012). Canadian Language Benchmarks: English as a second language for adults. Ottawa: Citizenship and Immigration Canada. Open access: http://www.language.ca

Donner, J. C. (2001). Using q-sorts in participatory processes: An introduction to the methodology. In Krueger et al. (2001: 24-59)

Douglas, D. (2000). Assessing Languages for Specific Purposes. Cambridge: Cambridge University Press.

Frentiu, L. & Gosa, C. (2011). English and the World of Work: Enhancing Career Opportunities through English Medium Exams. Romanian Journal of English Studies, 8, 101ff. Open access:
http://www.litere.uvt.ro/vechi/RJES/pdf/RJES-8-2011-complet.pdf#page=101

Kareva, V. & Dixon, H. (2009). The challenges of teaching English for Legal Studies in a European, non-EU country. SEEU [South East European University, Republic of Macedonia] REVIEW 5(1) 117ff Open access:
http://www.seeu.edu.mk

Krueger, R. A., Casey, M. A., Donner, J., Kirsch, S. & Maack, J.N. (2001). Social Development Papers Number 36: Social Analysis: Selected Tools and Techniques. Washington, DC: The World Bank, Social Development Department. Open access: http://siteresources.worldbank.org/INTCDD/Resources/SAtools.pdf

Labokaitė, A. & Sætre Ludvigsen, D. (2011). Legal English -- What is the Point? Studies About Languages 19, 5-14 [Lithuania]
Open access: http://www.erem.ktu.lt/index.php/KStud/article/view/940

Little, D. (2011). The European Language Portfolio: A guide to the planning, implementation and evaluation of whole-school projects. Graz: European Centre for Modern Languages, Council of Europe Publishing. Open access: http://www.ecml.at/tabid/277/PublicationID/65/Default.aspx

Luczak, A. (2011). Whose needs? Designing English for Legal Purposes Courses: A negotiation process. Open access: http://luczak.edu.pl/wp-content/uploads/2011/12/LUCZAK-Whose-needs.pdf

Ramlo, S. (2011). Using word clouds to present Q methodology data and findings. Human Subjectivity, 9(2), 99-111. Available from author http://drsueramlo.wikispaces.com

Ramlo, S. (2005). An application of Q methodology: Determining college faculty perspectives and consensus regarding the creation of a school of technology. Journal of Research in Education, 15, 52-69. Available from author http://drsueramlo.wikispaces.com

Shinebourne, P. (2009). Using Q Method in Qualitative Research. International Journal of Qualitative Methods, 8(1), 93-97 Open access: http://ejournals.library.ualberta.ca/index.php/IJQM/article/viewFile/980/5201

Sladoljev Agejev, T. & Pecotić Kaufman, J. (2009). Legal English in an Advanced Business English Course in Croatia: Identifying and Resolving Ambiguities. In Sočanac, Lelija; Goddard, Christopher; Kremer, Ludger (Ed.) (2009). Curriculum, Multilingualism and the Law. Zagreb: Nakladni zavod Globus (Biblioteka Jezik i pravo). pp. 407-426. Open access: http://web.efzg.hr/dok/PRA/jpecotic/sladolje%20agejev%20pecoti%20kaufman.pdf

Alma Ortiz

Universidad Nacional Autónoma de México, México City, México

alma.ortiz@cele.unam.mx

Proficiency Exams at CELE-UNAM: Guidelines for Analysis with the Common European Framework

Bio data

Alma Ortiz has a Masters in Applied Linguistics, full time teacher over more than 30 years in the Center of Foreign Language Teaching at UNAM-Mexico. Teaches Evaluation and Research Methodology in the Masters Program in Applied Linguistics; in the Teacher Training Program she is in charge of the Reading, Testing, and Practice courses offered to trainees of English. Her main academic interests are Evaluation, Reading Comprehension and Forensic Linguistics. She is the author of various articles on her main research areas.

Abstract

The international impact of the CEFR in assessing foreign languages is evident; the emphasis in certifying language proficiency for different purposes has been the cornerstone for exam users. Not only in Europe, but also around the world, diverse stakeholders were interested in analyzing the possibilities of adapting referents for their own academic or commercial purposes.

The Center of Foreign Languages of the National Autonomous University of México (CELE-UNAM) where university students certify their proficiency level in a foreign language has always being at the forefront of the teaching and testing trends. The Center has studied, analyzed, and questioned different approaches and tendencies to adapt or discard them to its students' needs.

This paper shares the results in the critical analysis of the parameters and referents published in the CEFR in order to confront them and establish the corresponding relationships with the exams produced in CELE.

The rigorous analysis included CEFR documents, ALTE information and CELE's own Framework of Reference; the proficiency exams used in CELE (English and French versions) were also analyzed in order to establish the guidelines for comparison, design and construction of new exams. The projects' products include information formats of exams, an operational definition of 'dominio' (the term commonly used to refer to the proficiency level of exam takers) and information for exam designers and exam users at all levels.

Short paper

The presentation shares the results of the critical analysis of the parameters and referents published in the CEFR in order to confront them and establish the corresponding relationships with the exams produced in the Center of Foreign Languages of the National Autonomous University of México (CELE-UNAM).

UNAM attends approximately 325,000 students per year:

Post graduates	26,169
Graduates	187,195
Bachillerato	110,119
Technicians (National School of Music)	930

From these numbers CELE, one of the certification language centers of the UNAM, certifies graduate and post graduate students. CELE, in its history, has studied, analyzed, and questioned different teaching and testing approaches and tendencies to adapt or discard them to comply with its students' needs; it has always being at the forefront of the teaching and testing trends, therefore compelling us to conduct the research with a formal and serious methodology.

Attending one of CELE's functions, through the Coordination of Evaluation (CEC) which impacts more than 100,000 students at graduate and post graduate students, one of the mandates in the current administration was mentioned as the "institutionalization and modernization of certification in foreign languages in the UNAM", transforming it as the main objective of this research.

The research team worked from the instruments in use: the bank of proficiency exams that the CEC offers to the test takers. The research was based on studying facts, conducting informative sessions and discussing the pertinence of the documents at hand. The research team is formed by six full time teachers of CELE.

We relied on the testing experience of other colleagues when we asked them to give a holistic impression of the exams in use, impression in terms of number of sections, testing techniques, overall instructions, without a specific format. Although a valid practice we realized we needed a guideline, a reference, not only for those teachers but for new exam designers or for administrative reports.

Finding a common parameter to analyze our exams, to be able to homologate future exams and compare similarities and differences with international exams was guided by our main reference: the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEF) first version, September 2003. As our framework will be not only be helpful to develop English exams, the team studied and translated the Forms A1 to A7 in Chapter 7 Guidelines for reporting in order to select, reject and adapt them to our context. One of the findings was realizing that little information and almost no feedback was offered to test takers before or after taking the exam. It was argued that these exams being certification instruments were not subject to being analyzed by the 'non - passing' test takers; nevertheless the realization helped us in finding how and when this information could be given.

The exams in question are proficiency exams called 'examen de dominio' that for most graduate students have to be taken as a graduation requirement, and for some post graduates it is an entrance requirement. Both instances have created controversy, when a specific major does not require their students to read, research and/or present findings in the foreign language, how are students expected to demonstrate its proficiency at the end of their studies? If a post graduate program is much demanded, one of the first sieves is to ask for a high score in a language exam, therefore the exams become high-stake exams.

Controversy alone, the very term 'Dominio del idioma' is a conflicting one in the very term used while describing language requirements in the study programs, therefore, some schools call it differently: 'plan global', 'cuatro habilidades', 'posesión de un idioma', 'inglés general'. We decided to offer our own definition combining and adapting different ones, including that which the CEFR offers, but the term remains because it is the current and legal term in the University´s legislature. Our definition tries to offer the

"what" and the "how" and it mentions different levels of proficiency by which it could be understood, no definition is found in UNAM documents.

> Un usuario domina un idioma cuando es capaz de desenvolverse apropiada y eficazmente en diferentes situaciones de comunicación de la vida diaria en los ámbitos académico, social y laboral. Esto implica:
>
> - Tener conocimiento de su sistema lingüístico y fonológico.
> - Tener conocimientos léxicos en relación a una gran variedad de temas.
> - Comprender y utilizar las reglas sociales del lenguaje y los diferentes tipos de discurso.
> - Expresarse con fluidez y ser capaz de interactuar en diferentes ámbitos sociales.
> - Poseer una conciencia lingüística que le permita autocorregirse y utilizar estrategias de compensación.
>
> Se pueden definir niveles de dominio de un idioma en términos de las habilidades y/o destrezas que requieren los usuarios para desempeñar diferentes tareas y realizar diversos tipos de interacción.

Following our discussion in this attempt to compare and align local parameters with internationally recognized standards for language assessment: CEFR documents, ALTE parameters, CELE's own framework of reference, we realized that trying to offer a solution to all interested participants was not easy. The stakeholders include the candidates, the test constructors, and the receiving institutions.

The research and the information which will be publish by CELE next year will hopefully help every level of administrative committees and testing teams to reach a common understanding in this function of testing.

The testing teams will profit from information regarding external contexts of use where test takers have to interact. The CEFR's descriptive categories, in terms of 'domains' did not convince us: personal, public, occupational and educational were not descriptive enough for our students, so we coined our own domain: 'academic', where instances of interaction are listed as well as lists of materials that are usually consulted by university students. Test designers can profit from this information for text selection parameters. There is also a document where we describe the sections, time, techniques and type and length of materials to be used in these instruments.

One of the challenges was the information that candidates could or could not have regarding the contents of the exam and the range in which their proficiency is measured. Although university students are informed and fairly sophisticated, sometimes their negative attitude towards being evaluated is very strong. The exam is a highly charged requirement where their future plans are sometimes crushed. We came agreed on giving them information in terms of levels of proficiency, skills tested, description of the exam in terms of length of written and listening texts, allotted time; this information will be given before they take the exam and can be referred to if they fail the exam. The information will be visually attractive visually to make it a friendly referent.

The research is now in its final stage and will be published to comply with its main objective. The CEFR has being useful as a 'referent', the information and all the hard work behind it is recognized and appreciated, nevertheless the specific situations in order to adapt the referents have to rule above everything. No testing policy has to be adopted without being critically analyzed in one specific context and honoring testing traditions.

References

Alderson, J.C. (2002). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies. Strasbourg: Alderson, J.C. (Ed.) Council of Europe Publishing.

Alderson, J.C. (2004). Quality Control in Evaluation and Assessment. British Council Hungary. Retrieved from http://www.examsreform.hu/Pages/Articles.html

Alderson, J.C. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework, Language Testing, 22, (3), 257-260.

Alderson, J.C. (2007). The CEFR and the Need for More Research, The Modern Language Journal, 91, (4), 659–663.

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006) Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project, Language Assessment Quarterly 3 (1), 3-30, retrieved from: http://www.informaworld.com/smpp/title~content=t775653669~db=all~tab=issueslist~branches=3 - v3

Aligning Tests with the CEFR: Reflections on using the Council of Europe´s draft Manual. Martyniuk, W. (ed.) (2010) Studies in Language Testing, 33. Cambridge: Cambridge University Press.

Bourguignon, C., Delahaye, P. & Vicher, A. (2005). L'évaluation de la compétence en langue: un objectif commun pour des publics différents, Etudes de Linguistique Appliquée 140 – 4 : 459-473.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press

Council of Europe. (2003). Relating Language Examinations to the CEFR. Manual; Preliminary Pilot Version, retrieved from: http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp#TopOfPage

Costa, M.E. (1996). Aportes de las ciencias del lenguaje para la consideración de la calidad en educación y su evaluación. Revista Iberoamericana de Educación, 10, 79-99.

Davidson, F. & Fulcher, G (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. Language Teaching, 40, 231-241.

Davies, A. (2003). Three heresies of language testing research. Language Testing, 20 (4), 355-368.

Farmer, F. (2006). Cultural Awareness and the Common European Framework. In Garrido, I & González, A. (Ed.) ANUPI's Conference Proceedings.

Figueras, N., North, B., Takala, S., Verhelst, N. & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: a manual Language Testing, 22 (3), 261-279.

Figueras, N. (2007). The CEFR, a Level for the Improvement of Language Professionals in Europe. The Modern Language Journal 91 (4), 673 – 675.

Hulstijn, J. H. (2007). The Shaky Ground beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. The Modern Language Journal 91 (4), 663-667.

Ljalikova, A. (2008). La valorisation de l'évaluation certificative en Didactique de Langues-Cultures Etrangères. Cla.univ-fcomte.fr; retrieved from http://ressources-cla.univ-fcomte.fr/gerflint/Baltique2/Evaluation.pdf

Ortiz, A., López, E., Mallén, M.T., Lusnia, K., Marrón, M.A. & Byer, B. (2012). Exámenes de dominio del CELE y el MCRE: análisis comparativo. Contijoch, M.C. & Lusnia, K. (Coord.) Investigación y enseñanza de lenguas: andanzas y reflexiones. (pp.183-199). México: CELE-UNAM.

Purpura, J. (2004). Assessing Grammar. Cambridge: Cambridge University Press.

Riba, P. (2008). La descripción y la especificación de los niveles del marco común europeo para los idiomas: de Threshold Level al Profile Deutsch, una experiencia compartida. Conferencia en la UNAM 24 enero, 2008.

Saville, N. (2005). Interview: An Interview with John Trim at 80, Language Assessment Quarterly, 2 (4), 263-288; retrieved from http://www.informaworld.com/smpp/title~content=t775653669~db=all~tab=isseslist~branches=2 - v2

Springer, C. (1999). Comment évaluer la compétence de communication dans le cadre d'une interaction spécifique : de quel type de critères pragmatiques avons-nous besoin?, Les Cahiers de l'Apliut, mars 1999.

Taylor, L. (2008). Guidelines for authors in writing and formatting typescripts. Studies in Language Testing, 10. Cambridge, Cambridge University Press.

Weir, C. (2005). Limitations of Common European Framework for developing comparable examinations and tests. Language Testing, 22 (3), 281-300.

Tina Rutar Leban, Ana Mlekuž, Karmen Pižorn & Tina Vršnik Perše

Educational Research Institute, Ljubljana, Slovenia
University of Ljubljana, Ljubljana, Slovenia

tina.rutar@pei.si - ana.mlekuz@pei.si - karmen.pizorn@pef.uni-lj.si - tina.vrsnik@pei.si

The Relation between Foreign Language Achievements of Slovenian students Included in ESLC and their Can-Do Statements

Bio data

Tina Rutar Leban holds a PhD in psychology. She graduated from the department of Psychology at the faculty of Arts of the University of Ljubljana in 2002. In the same year, she enrolled for post-graduate studies in psychology and finished her studies in 2011 with a doctoral dissertation entitled Subjective theories of pre-school teachers: Predictive value of personal characteristics of pre-school teachers and the perceived options of professional development. She works as a researcher at the Educational Research Institute in Ljubljana. The main fields of her research comprise evaluation studies in education, children's rights in education and teachers' subjective theories at pre-school level and above.

Abstract

The paper presents the results of the European Survey on Language Competences (ESLC) along with the links between 'can do' performance statements based on CEFR and areas of linguistic skills (listening, reading, writing) in Slovenia compared to other educational systems included in the survey. The survey was established to provide participating countries with comparative data on foreign language competence and insights into good practice in language learning. Sixteen European educational systems took part in the survey. Students (in the last year of lower secondary education ISCED2 or the second year of upper secondary education ISCED3) were tested in the two most widely taught foreign languages in their country chosen from the five tested languages: English, French, German, Italian and Spanish. Each sampled student was tested in one language only. The language tests covered Listening, Reading and Writing Language Skills and were based on CEFR. Each student was assessed in two of these three skills. Each student also completed a Questionnaire about his home and school environment regarding foreign languages. A representative sample of 1.500 students was chosen for each foreign language in each educational system. Altogether 53.000 students were tested together with 5.000 foreign language teachers and 2.500 school principals (In Slovenia 290 teachers and 163 school principals).

Students responded to 16 'can-do' statements, providing a self-evaluation of their competence in the tested language. The statements were taken directly or adapted from the descriptor used in the CEFR to illustrate the levels. Statements were chosen to be relevant to the target population. The paper presents the results of the 'can-do' statements, moreover it verifies the relationship between can-do self-ratings of students and their achievement at each of the tested skills in Slovenia and compares the results with other included educational systems.

Short paper

Introduction

Learning and teaching of foreign languages has become essential not only in countries with less widely spoken languages but also in those whose languages dominate and regulate science, research and international trade. As the British National Centre for Language (CILT)[1] states businesses that proactively use language skills achieve on average 45% more export sales. It has also been estimated that improving language skills could add up to £21 billion per year to the UK economy.

However, it is countries whose official languages belong to less widely spoken languages that appear to be fully aware of the importance of the foreign language skills of their citizens. They have started to implement different measures in the course of the compulsory education to ensure that their citizens become independent and active users of foreign languages (esp. English), regardless of their gender, education and ethnicity. The period of the compulsory education is also the time when a person develops empathy for other people, cultures, customs and religions. Learning foreign languages may, therefore, bring twofold benefits. On one hand, students are able to communicate in an additional language; on the other, they develop intercultural awareness (the ability to use the language in socially and culturally appropriate ways). Moreover, they may lose fear of being different and become more tolerant to the otherness and therefore enlightened citizens of the world.

The Common European Framework of Reference (CEFR)

The Common European Framework of Reference (CEFR)[2] for Languages as the basic document of European language policy provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks etc. across Europe. It describes in a comprehensive way what language learners have to learn in order to use a language for communication and what knowledge and skills they have to develop. The description also covers the cultural context in which language is set. The Framework defines levels of proficiency which allow learners' progress to be measured at each stage of learning and on a life-long basis. In the table below you can find the descriptions of each of the CEFR levels.

Proficient user	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.
	C1	Can express ideas fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.

[1] http://www.cilt.org.uk/home/about_us/a_new_agenda_for_languages/our_goals/languages_for_our_economy.aspx

[2] http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Independent user	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise while travelling in an area where the language is spoken. Can produce simple connected text on topics that are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic user	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Table 1: Description of CEFR levels

European Survey on Language Competences

The European Survey on Language Competences (ESLC) was established to provide participating countries with comparative data on foreign language competence and insights into good practice in language learning (European Commission, 2012a). The aim was also to contribute to the goals of the European Commission to improve the knowledge of fundamental language skills in Europe (2002) and the establishment of an indicator of language competence (2005). In 2007 the European Commission issued a document entitled the Framework for the European survey on language competences, which gave a more detailed description of the study and presented it to the European Council. In February 2008, the management of the study has been entrusted to a consortium named SurveyLang, which consisted of eight organizations: the Centre international d'études pédagogiques (CIEP) (language testing in French), Gallup Europe (computer software for testing and translation), Goethe-Institut (language testing in German), Instituto Cervantes (language testing in Spanish), CITO-National Institute of Measurement in Education (analysis and questionnaires), University of Cambridge ESOL (English language testing, organization, field operations), Universidad de Salamanca (language testing in Spanish), Università per Stranieri di Perugia (language testing in Italian). The management team of the survey regularly reported on the activities of the survey to the European Commission and the Advisory Board. The Advisory Board was composed by the representatives of the European countries.

The purpose of the article

The paper investigates the results of the European Survey on Language Competences (ESLC) along with the links between 'can do' performance statements based on CEFR and

areas of linguistic skills (listening, reading, writing) in Slovenia compared to other educational systems included in the survey. Moreover, it investigates the link between students' self-evaluation on can-do statements and their achievements according to CEFR levels in Slovenia. Students responded to 16 'can-do' statements, providing a self-evaluation of their competence in the tested language. The statements were taken directly or adapted from the descriptor used in the CEFR to illustrate the levels. Statements were chosen to be relevant to the target population. The paper explores the relationship between CEFR levels and students' self-evaluation with the help of can-do statements. The paper tries to investigate to what extend the students' answers to can-do statements correspond to CEFR levels and to what extend the students are realistic about their own skills. Furthermore, the paper presents the results of the 'can-do' statements. What is more, it verifies the relationship between can-do self-ratings of students and their achievement at each of the tested skills in Slovenia and compares the results with four selected educational systems. The educational systems were selected based on student's performance, therefore two educational systems with the poorest performance and two educational systems with the highest performance were included in the comparison.

Methodology
The Survey tested language competences in listening, reading and writing in the two most widely taught European languages (English, French, German, Italian and Spanish) in each country on a representative sample of students in the final year of compulsory education.

Sampling
A representative sample of 1500 students was chosen for each foreign language in each country. Altogether 53.000 students were tested in the first administration of ESLC in 2011. The participating students were attending the last year of lower secondary education ISCED2 or the second year of upper secondary education ISCED3 (European Commission, 2012b). The Survey also included 5.000 foreign language teachers and 2.500 school principals. They were administered an on-line Questionnaire about the school foreign language environment similar to the questionnaire for students.

The Survey included 14 European countries and 16 educational systems: Belgium with three different linguistic groups, Bulgaria, Croatia, England, Estonia, France, Greece, Malta, Netherlands, Poland, Portugal, Slovenia, Spain and Sweden.

Language tests
SurveyLang brought together five of the largest and most important organizations in the field of language testing (Cambridge ESOL, CIEP, Goethe-Institut, Instituto Cervantes, Universidad de Salamanca and CVCL Università per Stranieri di Perugia), in order to be able to develop quality testing instruments for the Survey. The language tests were based on the Common European Framework of Reference for languages (CEFR) but needed to be adjusted to the socio-cognitive characteristics of the tested population (14 and 15 year-olds). SurveyLang professionals determined the testing competences at each reference level of the CEFR (A1 to B2). Language tests consisted of three language skills: listening, reading and writing. Each student was tested in two of the above three language skills.

Questionnaires
In addition to language tests, the participants (students, teachers and school principals) completed a questionnaire. Each group of participants answered a different questionnaire but the purpose of all three questionnaires was to collect additional information about the language learning in the participating country.

In this paper the focus is on can-do statements which were a part of the Questionnaire. The students responded to 16 can-do statements (4 for each skill, speaking included)

and provided a self-evaluation of their competence in the tested language. The statements used in the Questionnaire are based on descriptor scales used in the CEFR to illustrate the levels. The statements are adapted or taken directly from the descriptor scales in CEFR, moreover they are applicable to the target population. The table below shows the can-do statements included in the Questionnaire. Each can-do statement corresponds to one CEFR level and one competence.

	Reading	Listening	Writing	Speaking
B2	I can scan quickly through long and complex texts, locating relevant details.	I can understand most TV news and current affairs programmes.	I can write clear, detailed descriptions, such as a review of a film, book or play.	I can explain my viewpoint on a topical issue giving the advantage and disadvantage of various options.
B1	I can recognise significant points in straightforward newspaper articles on familiar subjects.	I can understand the main point of radio news bulletins and simpler recorded material about familiar objects delivered relatively slowly and clearly.	I can write personal letters describing experiences, feelings and events in some detail.	I can enter unprepared into conversation and express personal opinions and exchange information on familiar topics.
A2	I can understand a letter from a friend expressing personal opinions, experiences and feelings.	I can understand what is said clearly, slowly and directly to me in simple everyday conversation, if the speaker can take the trouble.	I can write very short, basic descriptions of events, past activities and personal experiences.	I can tell a story or describe something in a simple list of points.
A1	I can get an idea of the content of simple informational material and descriptions, especially if there is visual support.	I can understand questions and instructions if people speak carefully and slowly, and I can follow short, simple directions.	I can write a few words and phrases that relate to myself, my family, where I live, my school.	I can ask and answer simple questions, make and respond to simple statements on very familiar topics.

Table 2: Can-do statements in the Student Questionnaire of ESLC

The CEFR has six proficiency levels, however in ESLC only 4 proficiency levels were used (levels from A1 to B2) due to the fact that students at the age when they were included in the survey cognitively are not yet able to reach the knowledge needed for users at levels C2 and C1. Moreover, the ESLC used another level named Pre-A1, which denotes the level of knowledge which does not meet the criteria of knowledge needed for level A1.

The main study

The main study was conducted in February and March 2011 on a sample of approximately 70 schools (1500 students) for each tested language in each participating country. The sample was determined to ensure the representativeness of the selected test population for each participating country.

Results and Interpretation

For the purpose of this paper, we present and discuss only a limited amount of data gathered and analysed in the ESLC survey.

Figures 1 and 2 below present percentage of students at each of the CEFR levels who endorsed a certain amount of can-do statements for English and German language separately. The purpose of this figure is to see whether students in Slovenia evaluate their own knowledge adequately according to their achievement on the language test (CEFR level they reached).

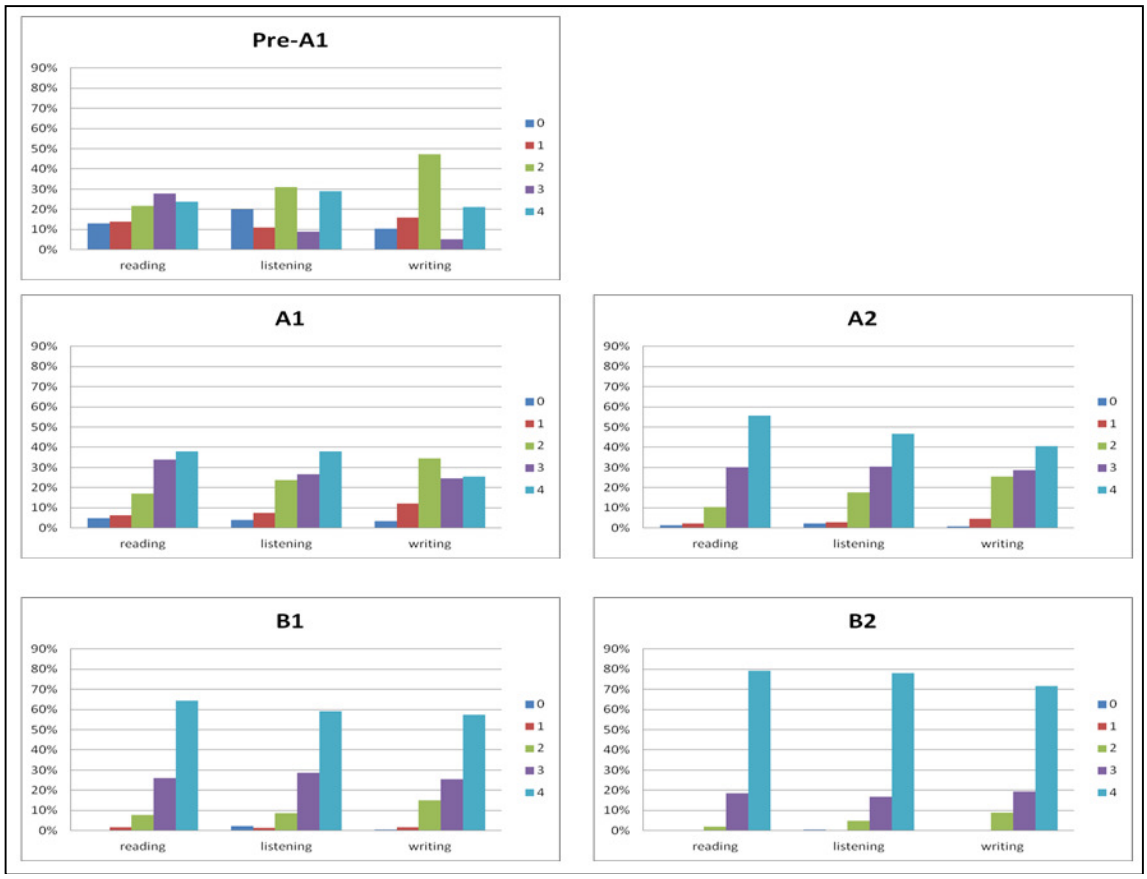


Figure 1: Can-do statements by skill and by CEFR level for Slovenian students (English)

Figure 1 shows that students overestimate their own competences, since the majority of the students endorsed 2 or more can-do statements at all CEFR levels. However there is a trend where students at Pre-A1 level endorsed less can-do statements than students at B2 level. The majority of students who did not endorse any of the can-do statements are according to their competences at Pre-A1 level, on the other hand the majority of students who endorsed all can-do statements reached the B2 level. Nevertheless, at level A1 the majority of students endorsed 2 or more can-do statements for each skill, at level A2 the majority endorsed 3 or more can-do statements and at level B1 the majority of students endorsed 4 can-do statements.

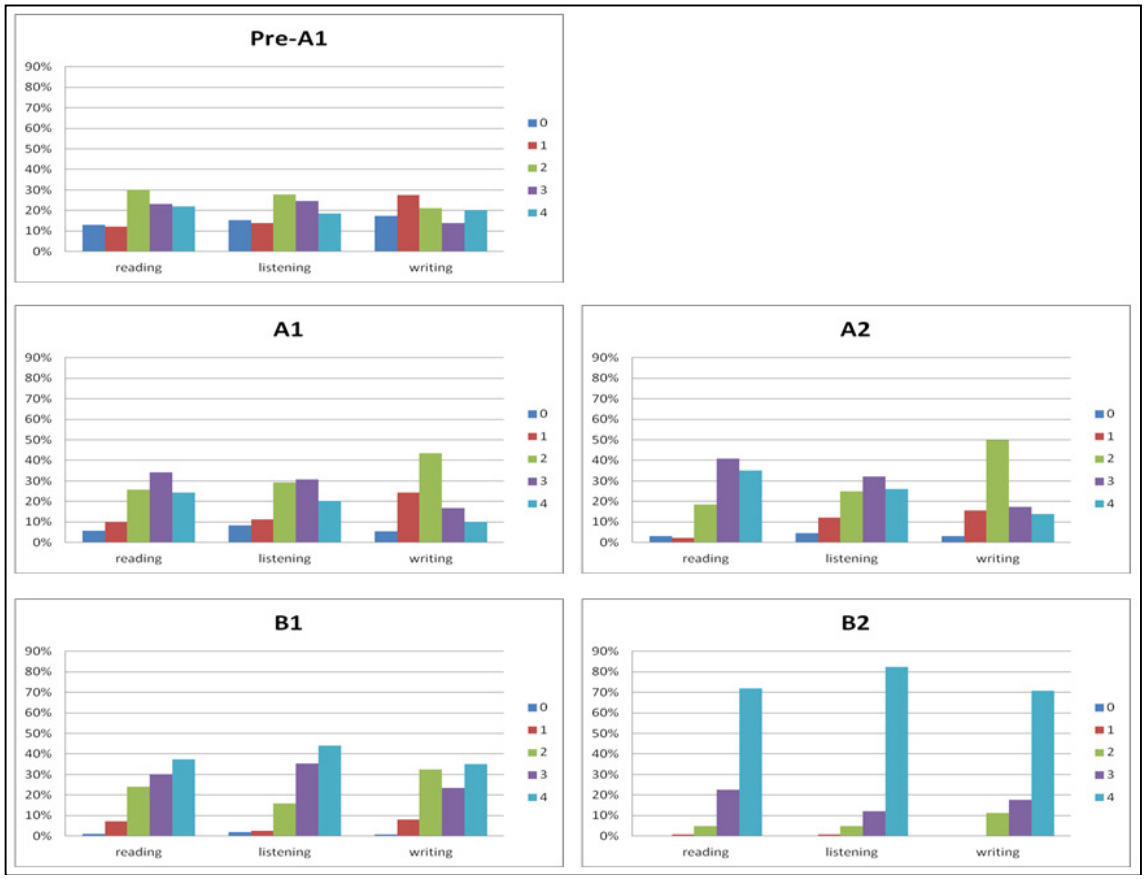


Figure 2: Can-do statements by skill and by CEFR level for Slovenian students (German)

Similarly, also Figure 2 for German language shows that students overestimate their own competences. What is more, also this figure shows a trend where students at Pre-A1 level endorsed less can-do statements than students at B2 level.

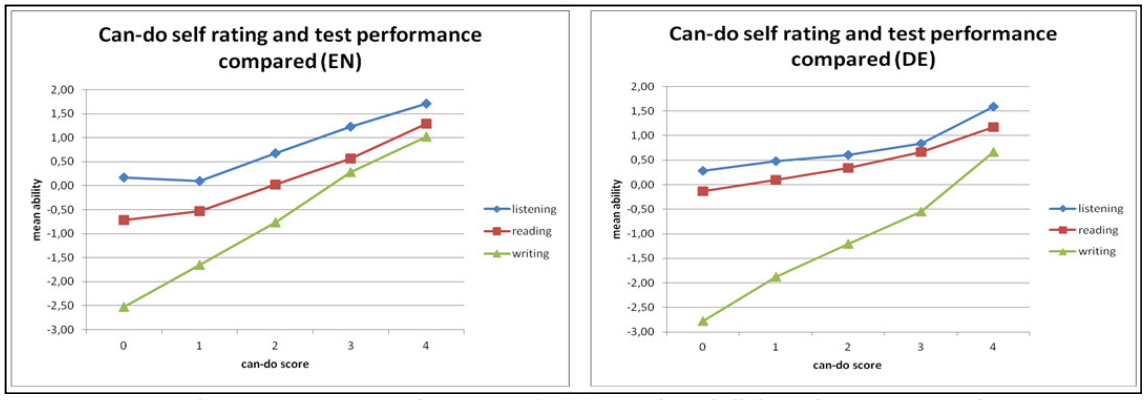


Figure 3: Can-do statements and test performance by skill for Slovenian students

Figure 3 shows a positive relationship between can-do statements and student performance for all skills. Students percive the writing as the most difficult, furthermore they feel that listening competence is the easiest. Students' test performance also reflect the latter perceptions, since students achived lower scores in writing and higher scores in listening. This is the case for both tested languages, however students achived lower scores in German language and hence they percive German language as more difficult than English.

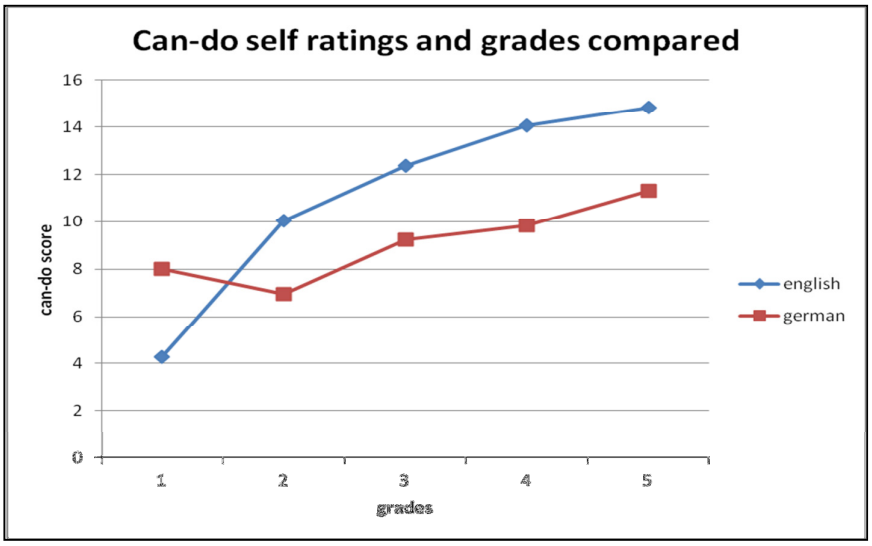


Figure 5: Can-do statements and students' grades for Slovenian students

Figure 5 also shows a positive relationship between can-do statements and final school grades for both languages. Students who endorsed more can-do statements also have higher final foreign language grade at school. Only one student who was tested in German language had final grade 1, therefore the german curve deviates from the expected curve.

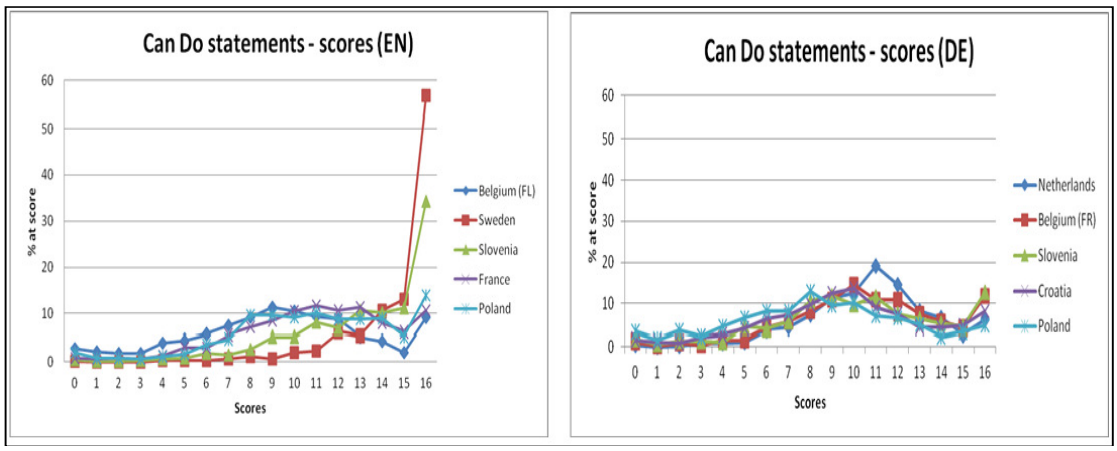


Figure 6: Can-do statements – comparison with chosen educational systems

Figure 6 shows the comparison of the percentage of students which endorsed each of the can-do statements from different educational systems. The educational systems were selected based on student's performance, therefore two educational systems with the poorest performance (France and Poland for English language and Croatia and Poland for German language) and two educational systems with the highest performance (Belgium (FL) and Sweden for English and Netherlands and Belgium (FR) for German) were included in the comparison. As it appeared, a proportion of students tested in English language had adopted a strategy of simply endorsing all the statements, all students with perfect scores were removed from the interpretation. As it is shown in the figure above students evaluate their own abilities approximately the same in all selected educational systems, even though their level of knowledge is significantly different. What is interesting is that Flemish students evaluate their own knowledge in English much lower than students from other selected educational systems, even though their overall achievement was higher than the majority of the educational systems included in ESLC.

Conclusion

The paper tries to analyze the relationship between the student knowledge achievement according to CEFR levels and their self-evaluation on can-do statements. One of the issues in the field of foreign language knowledge testing is, whether it is possible to comparably asses the level of knowledge according to CEFR levels and according to individual self-evaluation with the help of can-do statements. Analysis of the data shows that the relationship between the self-evaluation with the help of can-do statements and the knowledge test grade is positive[1], however the self-evaluation is not as accurate and sensitive as the evaluation with the knowledge test.

The paper shows that students when assessing their knowledge with self-evaluation tend to overestimate their own abilities. This is especially true for those students whose level of knowledge is lower. On the other hand, while analysing overall relationship between actual knowledge of students and students' self-evaluation of their knowledge, the correlation is positive regardless of which knowledge measure we use - CEFR level or school grade. Moreover, when comparing the results between selected countries, the data show, that students assess their own abilities more or less the same in all selected educational systems, even though their level of knowledge is significantly different.

References

European Commission. (2012a). First European Survey on Language Competences: Final Report. Brussels: European Commission.

European Commission. (2012b). First European Survey on Language Competences: Technical Report. Brussels: European Commission.

http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

[1] The correlations are: listening (EN) r = 0,40; reading (EN) r = 0,43; writing (EN) r = 0,42; listening (DE) r = 0,42; reading (DE) r = 0,35; writing (DE) r = 0,37. All correlations are statistically significant.

Cédric Sarré

Université Paris-Sorbonne, Paris, France

cedric.sarre@paris-sorbonne.fr

CLES, a Model Framework for 21st Century Higher Education Language Certification?

Bio data

Dr. Cédric Sarré is a Senior Lecturer in English as a Foreign Language (EFL) and English Language Teaching (ELT) at Université Paris-Sorbonne, IUFM (School of Education), France. His research interests include ESP course development in online settings, the integration of technology – especially CMC – in language learning and teaching, language proficiency testing and teacher education. He has been a member of the CLES national board since 2005.

Abstract

This paper aims at presenting the French Higher Education Language Certificate (CLES – Certificat de competences en Langues de l'Enseignement Supérieur), a task-based language assessment created in 2000 by the Ministry for Higher Education and based on the CEFR. It is currently available in 10 different languages and at three different levels (CLES 1, CLES 2 and CLES 3) corresponding to levels B1, B2 and C1 of the CEFR. The format of the test is that of a scenario which puts test takers in a realistic situation with a specific mission to complete, the completion of the mission requiring students to read texts, listen to documents, write and speak/interact. In addition to the fact that CLES is available to university students for free, it also has a number of advantages over other better-known private language tests which will be presented.

After a brief presentation of the context in which CLES was developed, the theoretical framework, test specifications and national organisation of the certification will be examined. Then, we will discuss the validity, reliability and feasibility of the test, along with aspects that could potentially be improved, some of which as a direct consequence of the CEFR descriptors. Finally, the positive washback effect of the test on language teaching at university level – which far outweighs any of the issues raised – will be dealt with.

Short paper

Introduction

Language teaching in French Higher Education for students specializing in subjects other than languages has always been very different from one university to the other - different contact times, different number of modules per degree, different ECTS credits attributed, etc. – mainly because the official recommendations from the Ministry for Higher Education have always been purposefully vague: the 2011 decree (which modified the 2002 original decree) setting out recommendations for Bachelor's degrees states that universities must give students the means to acquire language skills which will enable them to "read, write and speak in at least one foreign language" (Legifrance, 2011) and leaves it to the universities to organize the way they will provide training to reach these objectives as it doesn't give any detail. Interestingly, oral comprehension (listening) doesn't seem to be a skill worth developing in Higher Education as it isn't mentioned… As

for Master's degrees, they have to include training which will enable students to "validate their mastery of at least one foreign language" (Legifrance, 2002), but the official recommendations fail to mention how this validation should take place – with an exam, a language certification? – and at what level – as it isn't clear what "mastering" a language means… In spite of these very inexplicit recommendations which led to very different local organizations and, consequently, degrees, things have recently started to evolve towards common practices mainly thanks to the introduction of CLES (Certificat de competences en Langues de l'Enseignement Supérieur), the French Higher Education Language Certificate.

CLES was created in 2000 as a result of the Ministry's proactive policy consisting in promoting language learning in Higher Education and was considered to be a possible answer to several problems:

1. The language tests available at the time[1] were not considered to be accurate enough indicators of a student's ability to communicate in real life situations in a foreign language, as most of them did not directly assess productive skills (writing, speaking, interacting) but offered extrapolated correlations about the test-takers' proficiency level in the non-tested productive skills from their level in the tested receptive skills. These extrapolations were not satisfactory for many prospective employers who ended up hiring graduates with high test scores who could not use the language in everyday work situations. This has recently been confirmed by Liao et al.'s 2010 study, commissioned by ETS, whose objective was to find correlations between the proficiency levels in the TOEIC listening and reading test and those in the TOEIC speaking and writing test. Indeed, the study concluded that "distinct aspects of language proficiency (…) cannot be adequately assessed by other tests" (p.11), in other words, it is not possible to accurately extrapolate language proficiency levels in specific skills (writing or speaking, for example) from the assessment of other skills (reading or writing), contrary to what was claimed for almost 30 years.
2. The language proficiency tests available were mostly provided by private companies. Resorting to commercial tests in Higher Education was – and still is – somewhat ethically problematic for those teaching in public/state universities, as (i) teachers then become instrumental to the commercial success of these companies through the use of public resources (teacher time and facilities for the administration of the test) and as (ii) subcontracting is a way of acknowledging that language teachers in this sector are not capable of testing their students' proficiency level, which couldn't be further from the truth.
3. None of the tests available at the time made it possible to assess students' language proficiency in academic contexts using the same format for several languages.

The CLES experiment was first launched in 2000: the first experimental phase took place between 2000 and 2002 and was followed by a second experimental phase from 2003 to 2005 which involved over 8,000 test-takers. Then, in 2007, CLES was officially recognized as a viable and operational language certification that could be implemented nationwide through a new ministerial decree. CLES is the result of the fruitful collaboration between language teachers and researchers who believe in the need for and viability of a non-commercial public language certification in French Higher Education.

CLES description

Theoretical framework

As CLES was developed concomitantly with the publication of the Common European Framework of Reference for Languages (CEFR), it is grounded in the action-oriented

[1] The TOEIC, for example, rebranded "Listening and Reading", "Speaking and Writing" and "Four skills" in 2006, only existed in its basic version which did not assess productive skills.

approach: the basic principle of CLES papers is that they are presented in the form of a scenario, which means that all parts of the test are interlinked and that they all lead to the completion of a clearly stated mission within a specific context. This, of course, shows that CLES-takers are considered as "social agents, i.e. members of society who have tasks (not exclusively language-related) to accomplish in a given set of circumstances, in a specific environment and within a particular field of action" (CEFR, 2001, p.9). The scenario thus provides the "wider social context" (circumstances, environment and field of action) "which alone is able to give [tasks] their full meaning" (CEFR, 2001, p.9), as well as the realistic mission (macro-task) to accomplish. The action-oriented approach, when applied to assessment, implicitly refers to Task-Based Language Assessment (TBLA) which consists in "evaluating, in relation to a set of explicitly stated criteria, the quality of communicative performances elicited from learners as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge" (Brindley, 1994, p.74). Accordingly, assessment tasks are viewed as "devices for eliciting and evaluating communicative performances from learners in the context of language use that is meaning-focused and directed towards some specific goal" (Ellis, 2003, p.279). The basic assumption of TBLA is that linguistic competence (knowledge of vocabulary and grammar) is not enough to be able to achieve goals in real-life social situations as communicative competence also includes sociolinguistic, pragmatic and discursive skills (Mislevy et al. 2002). These are elements which are taken into account in the language performance assessment of CLES-takers.

Test specifications

CLES scenarios start with the situation description, that is the realistic context that sets the stage for language use throughout the test and that explicitly mentions the role that test-takers have to take on, as well as the mission they have to complete. In order to complete their mission, which takes the form of written and oral language products, they are required to read and watch/listen to a set of authentic documents (text, video and/or audio files). The oral and written comprehension parts of the test enable test-takers to notice and extract from the documents essential content (ideas and lexis/structures) that will be useful for the completion of their mission: the comprehension documents thus serve as a source of input, just as the comprehension tasks serve the production tasks.

CLES allows students to certify their language proficiency at three different levels: CLES 1 (level B1), CLES 2 (level B2) and CLES 3 (level C1). CLES's relative youth, which is sometimes put forward as a drawback by its opponents, turns out to be a real upside as it stems from the CEFR and the Council of Europe's earlier publications: not only are the CLES proficiency levels based on the CEFR's common reference levels – unlike most language tests which had to subsequently try to find correlations between their own scoring systems and the CEFR's levels – but the marking schemes used for the production parts of the test are also based on the CEFR's language proficiency descriptors for each of the three levels. As for the contexts and themes chosen to develop CLES scenarios, they also have their source in the CEFR and the contexts of language use described in the global scale (CEFR, 2001, p.24): as level B1 is associated with "familiar matters regularly encountered in work, school, leisure" and "situations likely to arise whilst travelling in an area where the language is spoken", CLES 1 scenarios deal with situations connected to living and studying abroad; since level B2 mentions "complex texts on both concrete and abstract topics", "a wide range of subjects" and the ability to "explain a viewpoint on a topical issue giving the advantages and disadvantages of various options", CLES 2 scenarios deal with general topical issues, within the context of Higher Education, presented from different perspectives; finally, level C1 mentions "a wide range of demanding, longer texts" with "implicit meaning", "complex subjects" and the ability to use the language "for social, academic and professional purposes", which explains why CLES 3 scenarios deal with topics within the students' specialist field presented from an academic perspective.

In terms of skills tested, CLES 1 assesses language proficiency in reading, listening, writing and speaking, while CLES 2 assesses reading, listening, writing and interacting, and CLES 3 assesses all five skills. CLES is thus a complete language certification as it directly assesses all five skills. However, CLES is a new kind of language test as the assessment of receptive skills is item-based at levels B1 and B2 (not at level C1), whereas the assessment of productive skills is task-based (at all three levels). In addition, CLES is a multilingual certification: students' language proficiency can be assessed in 10 different languages (English, German, Spanish, Portuguese, Italian, Arab, Greek, Russian, Polish and Chinese) using the same paper format and topics regardless of the language. CLES is thus a tool that can play a part in the assessment of students' "plurilingual competence", as advocated by the authors of the CEFR (2001, p.133).

National organization

CLES's non-commercial business model rests on a national network of over 50 accredited universities and Higher Education institutions, organized in nine regional groups: after their accreditation by the Ministry for Higher Education, universities receive extra funding for the implementation of CLES, as the objective is to make sure that CLES is offered to students free of charge.

The CLES network is coordinated by a national board comprising 12 members and is in charge of the entire certification process, from paper writing to paper rating, as the pooling of strengths and resources is at the heart of the process. As soon as a university becomes an accredited CLES centre, they enter the network and commit themselves to contributing to the running of the certification at the national level, starting with paper writing: test papers are (a) written by local teams following a strict test specification document (and after attending specific training), then (b) evaluated by a pair of experimented paper writers who make recommendations and ask for adjustments, and finally (c) validated by the national paper validation committee. Once validated, papers become part of the national bank and can be used in any CLES centre. CLES paper writers are paid by their own institution, the rule being that a CLES centre has to write one paper for every five certification sessions organized, which means that these costs can be integrated in local CLES budgets from the start.

Whenever a CLES centre wishes to organize a certification session, they have to "place an order" with the national coordinator who checks that they are accredited, that they have produced – or are in the process of producing – the correct number of papers, and then gives them access to a paper from the national bank (preferably not a paper written by the requesting centre). This just shows to what extent inter-institutional collaboration is at the basis of the CLES organization. Each member of the CLES network is bound to the others by a moral and professional commitment, not by commercial ties.

Discussion

Validity and reliability

If CLES seems to possess the three primary qualities of a communicative language test as defined by Fulcher (2000) (as it involves performance, has an authentic communicative purpose and is scored on real-life outcomes), the issues that need discussing here are those traditionally associated to language assessment (CEFR, 2001, p.177), namely validity and reliability.

Validity first concerns test construct, i.e. the fact that "what is actually assessed (the construct) is what, in the context concerned, should be assessed" (CEFR, 2001, p.177). Given that CLES rests on the notion of communicative competence (see section II.1.), the four components of communicative competence are taken into consideration in the assessment, as illustrated in the CLES production assessment grids which clearly include linguistic, discursive, sociolinguistic and pragmatic descriptors. Besides, CLES aims at assessing language proficiency in academic contexts, which is exactly what it does given the scenario topics chosen and the true-to-life situations test-takers are put in (see

section II.2.). In this respect, CLES seems to be "a way of achieving a close correlation between the test performance, i.e. what the testee does during the test, and the criterion performance, i.e. what the testee has to do in the real world, and thus of ensuring the validity of the assessment" (Ellis, 2003, p.279). Another aspect of test validity is that of face validity, that is "the extent to which the test is perceived as acceptable by stakeholders, including testees" (Ellis, 2003, p.282): on this particular point, feedback on CLES couldn't be better as both stakeholders and students recognize the relevance of the test's format and content which are viewed as closer to real world situations than other tests, and, consequently, as a more accurate way of assessing language proficiency. However, CLES suffers from its youth and relative lack of recognition: its face validity isn't as good among prospective employers and higher education institutions outside France, simply because it isn't well-known.

When it comes to test reliability, CLES is nothing like psychometric tests made up of closed questions which prioritize reliability and objectivity by using statistical procedures extensively. Consequently, reliability is a key issue for CLES, as it is for any type of assessment, as it deals with "the extent to which a test measures a candidate's proficiency in an error-free manner" (Ellis, 2003, p.310). In other words, what needs to be ensured is that a repeat test or a second rating would give the same result (the same measure of proficiency). As CLES, in the TBLA tradition, requires test-takers to produce language with a specific objective and in a particular social context and from which proficiency is measured, reliability mainly depends on "the accuracy of decisions made in relation to a standard" (CEFR, 2001, p.177) as human judgment is involved since raters have to make binary decisions (pass/fail) for each of the descriptors included on the evaluation grids. The problem here is clearly to limit rater subjectivity in order to approach objectivity, which can be done by providing "a rating scale, set of task requirements and marking criteria" (Milanovic, 2002, p.32), as well as "a brief description (…) of a typical performance" (ibid. p.33). Another idea is to "accompany descriptors of performance with actual examples of candidates' work" (ibid.). Ellis goes further as he sets out a list of four possible solutions to enhance TBLA's reliability (2003, p.311): making the test longer (to ensure provision of larger samples of language use), using two raters, training raters, and adjusting test scores thanks to statistical analysis. CLES has addressed the issue of reliability through several measures:

1. there are two versions of the evaluation grids – one simple one, and one which includes sample productions considered acceptable;
2. the answer key includes elements expected from a typical acceptable performance;
3. all raters have to follow specific training which aims at setting standards for the different task types (writing, speaking, interacting);
4. for the assessment of the oral interaction part of CLES 2, two raters are recommended.

Room for improvement

Although the CEFR's input has been invaluable in the development of CLES, its link to the CEFR is also a source for potential problems:

1. the CEFR descriptors sometimes lack elaboration, especially when it comes to linguistic competence associated to each level;
2. it seems difficult to establish clear cut-off points between the different levels when it comes to linguistic competence;
3. the multiplicity of scales for a given language activity in the CEFR also makes it difficult to get a global view of what a user/learner can do at a given level (in spite of the general scale).

This can be explained by the fact that the CEFR is "a point of reference, not a practical assessment tool" (CEFR, 2001, p.178), and it can also be considered as a work in

progress, as it states that there are "gaps in the descriptors provided", and that, for specific areas, "descriptors could presumably be written, but haven't been" (CEFR, 2001, p.37).
Another issue, although not directly related to the CEFR, is that of the conception of the test itself: CLES is a horizontal test, as it assesses language proficiency in the tested skills at a given level, unlike vertical tests which assess the test-takers' level on a continuum and grants them different proficiency levels for the different skills tested. The direct consequence is that test-takers need a pass for all the tested skills to get their certificate (a test-taker who has performed at level B2 in reading, listening and writing, but at B1 only in speaking won't be awarded the CLES 2 certificate). This obviously has a strong impact on CLES pass rates which are about 40% overall.

Conclusion
With 400 CLES sessions organized in 50 CLES-accredited centres in 2011 and a total of 35,000 test-takers, the CLES dynamics seems to attract growing interest from both students and language professionals who, for the first time, feel that they are an important link in the certification chain and that they can make a difference. Indeed, as test development is an iterative process, CLES is constantly evolving as more data becomes available and as further adjustments thus appear necessary.

In spite of the issues raised (especially as regards test reliability and reliability of the CEFR descriptors), CLES's gain in validity and its tremendously positive washback effect on language teaching methods far outweigh its drawbacks, if we bear in mind the fact that examinations always have a direct or indirect effect on teaching methods and that this effect can be either positive or negative (Heaton, 1990). Indeed, in the past, language professionals in French Higher Education could be trapped in an exam preparation cycle which involved doing a lot of past papers for certain psychometric tests which were sometimes made a requirement for students to graduate. In this case, the testing procedure (psychometric test) had negative effects on classroom practice (test preparation through extensive use of past papers). CLES's washback effect, on the contrary, is very positive as (1) it has contributed to the introduction of the CEFR in Higher Education, (2) it has encouraged language professionals and institutions to offer language courses with clearer objectives expressed in terms of proficiency levels, (3) it has had an impact on classroom practice which now includes the implementation, more often than not, of the action-oriented approach and task-based language teaching (TBLT). For all these reasons, it is strongly believed that CLES could provide a model framework for 21st century Higher Education language certification development and increase its European dimension by developing partnerships with institutions outside France.

Still, more research is needed to seek satisfactory solutions to the problems associated with TBLA in general (Ellis, 2003, p.311) and CLES in particular. This is one of the missions of the CLES scientific committee whose task will be to initiate and support various research projects in years to come. In this respect, a new framework of reference for language testing could be very useful.

References

Brindley, G. (1994). Task-centered assessment in language learning : The promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allsion & A. McNeill (eds), Language and learning, (pp.73-94). Hong Kong: Hong Kong institute of language in education, Hong Kong education department.

Council of Europe. (2001). Common European Framework of Reference for Languages.

Ellis, R. (2003). Task-based language learning and teaching. Oxford: Oxford University Press.

Fulcher, G. (2000). The "communicative" legacy in language testing. System, 28, 483-497.

Heaton, J.D. (1990). Writing English language tests. London: Longman.

Legifrance. (2002). Arrêté du 25 avril 2002 relatif au diplôme national de master. Retrieved from http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000771847&dateTexte=&categorieLien=id

Legifrance. (2011). Arrêté du 1er août 2011 relatif à la licence. Retrieved from http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000024457754&dateTexte=&categorieLien=id

Liao, C., Qu, Y. & Morgan, R. (2010). The relationships of test scores measured by the TOEIC listening and reading test and TOEIC speaking and writing tests. TOEIC Compendium. Retrieved from http://www.ets.org/Media/Research/pdf/TC-10-13.pdf

Milanovic, M. (2002). Language examining and test development. Strasbourg: Language policy division, Council of Europe. Retrieved from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/Language_examining_EN.pdf

Mislevy, R., Steinberg, L. S. & Almond, R. G. (2002). Design and analysis in task-based language assessment. Language Testing, 19 (4), pp.477-496.

Skehan, P. (1998). A cognitive approach to language learning. Oxford: Oxford University Press.

Susan Sheehan

British Council, Manchester, United Kingdom

susan.sheehan@britishcouncil.org

A Core Curriculum Inventory for General English

Bio data

Susan Sheehan is Adviser Learning and Teaching for the British Council. Her areas of specialism are testing and the CEFR. Susan has delivered courses on testing and assessment in many countries including Uruguay and China. She managed the development of the new British Council placement test. She is the editor of the Research Papers series: http://englishagenda.britishcouncil.org/research-papers and co-ordinates the English Language Teaching Research Partnership Awards.

Abstract

What should I teach and assess from A1? When should I introduce the present perfect and when should I test it? How can I improve classroom assessment? The British Council and EAQUALS joined together to answer these questions by creating a core curriculum inventory for CEFR levels A1 to C1 for English. It includes grammar, lexis, functions and notions and topic areas. Alongside the curriculum a series of tasks have been developed for teachers to use for classroom assessment.

My talk discusses the British Council/EAQUALS (The European Association for Quality Language Services) core curriculum inventory. I will begin with a description of the curriculum and the methodology used to create it. Then, I will describe the status of the curriculum and how it can be used by English language practitioners. My talk will conclude with recommendations for practice and areas for further research. Below, I give an outline of each stage of the talk.

The core curriculum inventory represents the core of English language taught at Common European Framework of Reference (CEFR) levels A1 to C1. It includes grammar, lexis, discourse markers, functions and notions.

A number of sources were drawn on, including an analysis of the language implied by CEFR descriptors, an analysis of content common to CEFR-based language school syllabi and popular coursebooks, and a teacher survey.

Each language point appears at the level(s) at which it is considered of most relevance to the learner in the class room. The inventory is a documentation of current best practice that can be used in conjunction with databases of learner language like the forthcoming English Profile. The core curriculum will assist teachers in planning classroom assessment. It has been used to inform the development of language tests.

Short paper

The CEFR is a fine document but what does it mean for the classroom? Many teachers have asked this question since the publication of the "blue book". In one sense, as one of aims of the original project to develop a system through which examinations of European languages could be compared, a possible answer would seem to be that the CEFR and the classroom are not closely linked. However, this does not take into account

the primary aim of the CEFR: to encourage reflection on current practice in relation to the specification of what is taught and the assessment of the successful learning of that content. In fact, the CEFR is relevant to all aspects of English language teaching. The British Council – EAQUALS core inventory of general English helps practitioners, both students and teachers, to make the CEFR relevant to the classroom. The abstract becomes concrete and practitioners can clearly see the language points including grammar and lexis, discourse markers, and functions and notions which make up the core of each CEFR level.

The initial data collection and analysis drew on a variety of sources. These included teacher surveys, a coursebook survey, and syllabuses from centres of excellence in implementing the CEFR into courses. So, the inventory is not telling teachers what to teach, rather it is describing what teachers are teaching with the intention of informing discussion and providing teachers and syllabus writers with support and guidance.
The inventory represents the core of English to be taught at CEFR levels A1 to C1. As the name suggests, it is the core, it is not the whole. Teachers and syllabus writers will define the total content of a course. The inventory provides guidance and support for those who are involved in course design. It provides the foundation for courses for institutions which aspire to reflect the aims of the CEFR in their course aims. The inventory documents one approach to realising an "action-orientated" approach to language learning and language use described in the CEFR. It is only one possible to approach to achieving such an aim and institutions could adapt our methodology to make an inventory for their own situation.

As outlined above the levels of the CEFR can be made explicit in terms of required vocabulary and grammar. Whilst the work described above is only one possible approach to the problem it does demonstrate that it is possible to make levels explicit. The levels should be made explicit as this information, it can be argued, is of use to everyone involved in the learning, teaching and assessing of English. In particular equipped with this type of knowledge learners are better able to judge the quality of a course and the materials provided and where best to direct their efforts.

The "action-orientated" approach to language described above is operationalised through the Scenarios. These are tasks and criteria which can be used for teacher, peer- and self-assessment. The components of a scenario include can-do statements, criteria and an elaboration of language competences. The competences are strategic, pragmatic (functional, discourse) and linguistic (grammatical, lexical, phonological). Scenarios can be used for both teaching and for different forms of assessment. A scenario is a holistic setting that: "encourages the integration of different aspects of competence in real (fistic) language use." Thus, the levels of the CEFR can be made explicit for learners and by doing so the connections between classroom language and language use in the real world are reinforced.

References

Ortega, A. & Sheehan, S. (2010). British Council – EAQUALS Core Inventory for General English Brian North. London: British Council.

Carol Spöttl & Kathrin Eberharter

University of Innsbruck, Innsbruck, Austria
Austrian Federal Institute for Education Research (BIFIE), Vienna, Austria

c_spoettl@hotmail.com - kathrin.eberharter@uibk.ac.at

CEFR Performance Descriptors and the Missing Formulae

Bio data

Carol Spöttl teaches at the University of Innsbruck, is a member of IMoF (Innsbrucker Modell der Fremdsprachendidaktik) didactic team and is currently setting up a research group within the newly established School of Education. Since 2007 she has been project leader of a government awarded research project to develop and deliver a new CEFR linked school leaving exam for English, French, Italian and Spanish. In addition to task development and item moderation, this work has involved extensive training of item writers, raters, and stakeholders for standard setting sessions.

Kathrin Eberharter graduated in English and German at the University of Innsbruck with a testing related dissertation and has an MA in Language Testing from the University of Lancaster. She is currently working for the Austrian Federal Institute of Education Research (BIFE) where she is an item moderator for the English writing and Language in Use sections of a newly implemented national exit-level examination. Furthermore, she is involved in teacher training on test task design and the rating of writing performances.

Abstract

The linguistic model underlying the CEFR is a functional notional approach additionally employing descriptive theory to scale what a language learner/user can do with a language and with which assessors can assess performance and competence levels. However, the CEFR 'can do' performance statements refer to conspicuous searching for words (General linguistic range), incorrect word choice (Vocabulary control) or lexical gaps (Vocabulary range), implying an approach that focuses on isolated items of language and regards grammar and vocabulary as dichotomous entities rather than an approach to language description that has been the subject of more recent studies in the field. At present, descriptors beyond the basic user level (isolated phrases A1 VR; memorised phrases and formulae A2 GLR) lack any reference to multiword items (Moon 1997), formulaic language (Wray, 2002) or formulaic sequences (Schmitt 2004). The ubiquitous nature of these lexical items has been estimated to cover over 52% of written English and over 58% of spoken language (Erman & Warren, 2000) yet the performance statements as they stand do not cater for this frequency and prevalence which Wray views as a dynamic response to the demands of language use. She argues that, "recognizing the role of formulaicity is fundamental to the understanding of the freedom and constraints of language as a formal and functional system," (2002.5).

This practice-related paper will outline how CEFR performance descriptors have been used to support the task development of a national school leaving exam across the first and second foreign languages. It will describe how and where the descriptors have been constructive in guiding training and task development across three languages and outline attempts and the rationale behind incorporating formulaicity into both test construct and task development. Results from trials and live data will be presented.

Short paper

CEFR performance descriptors and the missing formulae
a) The description of the current practice

The research angle this presentation takes is threefold:

1. CEFR grammar and vocab descriptors have indeed been a support in task development by providing a common underlying construct across the languages (E/F/I/S) that we are currently working on. The descriptors have also proved instrumental for the development of a CEFR-linked rating scale for writing.
2. The performance descriptors were likewise useful in providing a common "test talk" with which to compare and discuss items targeted, item difficulty and item statistics across languages and trial populations. However, this process has also revealed limitations that lead us to believe the levels should be made more explicit in terms of required vocabulary and grammar.
3. The main innovative claim behind this work is the attempt to include Sinclair's idiom principle to task development in language in use and CEFR related rating of writing performances.

Current research and practice
Following the introduction of a CEFR linked secondary school curriculum in 2004, Austria has been developing and implementing a standardized national exit-level examination since 2007. For the first and second foreign languages the test consists of four CEFR-linked sections: Reading, Listening, Writing and Language in Use. The new exam laws have set the exit level for the L2 at B2 and L3 and 4 at B1.
This paper will outline how CEFR descriptors have been used to support the task development of this national school leaving exam across the first and second foreign languages. It will describe how and where the descriptors have been constructive in guiding training and task development across languages and outline attempts and the rationale behind incorporating formulaicity into both test construct and task development.

Operationalising CEFR descriptors in the assessment of Writing and Language in Use:
Descriptors in an analytic writing scale
Two CEFR-linked assessment scales, B1 and B2, were developed as a first measure to standardise the assessment of the writing performances. The analytic scales contain four independent and equally weighted criteria:

1. task achievement,
2. organisation and layout,
3. lexical and structural range, and
4. lexical and structural accuracy.

The starting point in scale development was the selection of the criteria; range, accuracy and coherence and cohesion from the Manual for Relating Exams to the CEFR (Council of Europe, 2009, table C4). It was also decided to not perpetuate the dichotomy of lexis and grammar found in other scales, but to amalgamate grammar and vocabulary and create two criteria called lexical and structural range (LSR) and lexical and structural accuracy (LSA). The criterion LSR looks at the range of structures and lexical phrases the test takers use for the set task and how appropriately they adapt their register to the set task. It thus allows for some words to behave more lexically in some contexts and grammatically in others. The criterion LSA looks at how accurately the test takers use structures and lexical phrases and takes into account accuracy, appropriateness, spelling and punctuation of the language produced. The descriptors on which these two criteria and their bands were built upon were taken from several different CEFR scales. For LSR

some of the descriptors were taken from the scales describing General Linguistic Range, Socio-linguistic Appropriateness, and Vocabulary Range. For LSA descriptors from the scales Grammatical Accuracy and Vocabulary Control were used.
Test specifications

Both the CEFR and the Manual for Relating Exams to the CEFR (Council of Europe 2001 & 2009), were of value in drafting assessment scales and test specifications for reading, listening, writing and speaking. But, for several reasons, they proved problematic and restricting in the development of specifications for Language in Use test formats. First, the sheer number of descriptors in the scales and subscales for writing far exceeds those for grammar and vocabulary. Also, although the relevant descriptors could easily be adapted for the writing scales, this was not found to be the case for the test methods required for the language in use. Here the performance descriptors did not appear to describe the competences observable in a typical language use test with closed and open response formats. Second, the CEFR descriptors do not provide the necessary detail to facilitate communication between item moderators, item writers and test users. For these reasons we decided to include the descriptors for grammar and vocabulary from the DIALANG project (Alderson, 2005).

Item development and item writer training
In the training of item writers to develop Language in Use tasks, the CEFR and DIALANG descriptors were of limited use. They were constructive in providing a common basis for all languages, but item writers themselves expressed the need for more explicit and at the same time less restrictive descriptors for their language in order to target linguistic features in a text more accurately. Different linguistic areas that could/should be targeted at a given level were identified for different languages. Some of the areas of discussion were (1) the importance of accuracy of spelling in the Romance languages; if the ability to place an accent correctly was viewed as part of word knowledge then the facility values in trial data plummeted (2) concepts that do not apply to all languages (e.g. phrasal verbs), or (3) linguistic features that were argued to appear and be required earlier in some languages than in others (e.g. the passive in Romance languages). To compensate for the lack of explicitness in some CEFR descriptors, insights from corpus linguistics proved useful in providing some guidance as to which elements in a text were suitable for test tasks on level B1 or B2 respectively. Corpora were therefore consulted to identify linguistic features that are more probable and, thus, natural (Purpura, 2004, p. 14). Frequency lists (e.g. Cobb; Davies, 2006; Lonsdale & Le Bras, 2009) are consulted in order to check for "common words", "less common expressions", "very frequent words" or "frequent collocations" for each language.

Incorporating formulaicity in the test construct and task design
To account for formulaicity and collocation – a crucial part of our test construct–, the test specifications of the Language in Use section were extended to include the following adaptations of existing descriptors at the appropriate levels:

- Can identify the appropriate collocational use for a given context
- Can complete a range of high frequent formulaic sequences to his/her field and most general topics
- Good command of most adverbs and adjectives and their collocational use

However, the paradigm change from the individual word approach in task development to a phrasal approach presented a further unexpected problem. It was a challenge at first for many item writers to recognize formulaicity in texts. They tended to target the small linguistic units as an item rather than seeing the unit as part of a larger chunk. Thus for example the preposition in would be targeted, when it was actually part of a larger chunk of language (e.g. to keep something in mind). To identify suitable target words or phrases, a method of textmapping (Sarig, 1989) was introduced and adapted to help identify level specific linguistic features, including lexical phrases, and ensure a common

understanding among item writers of what was indeed a phrase. At present three test methods have proved suitable in incorporating formulaicity in the Language in Use section of the exam: open gap-fill, banked gap-fill, and multiple choice gap-fill.

b) the discussion of one of the conference topics - Competence and Performance: can-dos and linguistic knowledge

The CEFR and linguistic competence

The CEFR describes three components of communicative competence: linguistic, socio-linguistic, and pragmatic competencies. Linguistic competence is further described in a range of scales and subscales amongst which are general linguistic range, vocabulary range, vocabulary control, and grammatical accuracy. The can-do statements describe how different levels of ability are expressed when a user productively uses the language. These statements, however, are formulated generally without any further specification of precise linguistic features to be expected of learners at a given level.

Further, as the CEFR is neutral with respect to language, this limitation is compounded when developing tasks across languages. What is a basic vocabulary in English and is this similar to French, Italian and Spanish? Where are the cut-off points between a basic or elementary vocabulary, sufficient vocabulary and a good range of vocabulary? Granted, research in the field has shown it is possible to identify frequency levels for vocabulary and this across some languages (Laufer & Nation, 1995; Cobb; Nation & Beglar, 2007) but to inform test development and make target learning goals more transparent and tangible, there is a clear need for vocabulary levels to be made more explicit.

An additional issue merits attention with the CEFR 'can do' performance statements and areas of linguistic knowledge. The CEFR tenders a view of language as a dichotomy with vocabulary at one end of the spectrum and grammar at the other. Areas such as vocabulary range and vocabulary control and grammatical accuracy are related to the accepted linguistic understanding of the time. Any reference to a phraseological approach is restricted to the A levels:

- Can write short, simple formulaic notes relating to matters in areas of immediate need.
- Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people, what they do, places, possessions etc.
- Has a limited repertoire of short memorised phrases covering predictable survival situations;

The question further posed by this research is whether a grammar vocabulary dichotomy is not only impractical but inaccurate given the latest focus and developments in language description. More recent research perspectives of language see grammar and vocabulary as two related and sometimes even interchangeable components of language. Larsen-Freeman and DeCarrico (2010) point out that it is "very difficult to isolate grammar and lexis into completely separate categories, because grammar does not exist on its own". They see it is "interdependent with lexis and, in many cases, grammatical regularity and acceptability are conditioned by words" (p. 25). The research presented here adopts Halliday's "lexicogrammar" approach (1994) including both categories but viewing them as "merely different ends of the same continuum – they are the same phenomenon as seen from opposite perspectives" (p. 15).

Willis (2003) reflects our experience in task development when he accurately states that the lexical phrase is not easy to define or grasp (p. 142). Skehan (1992) refers to lexical phrases as "ready-made elements and chunks" (p. 186). Other researchers prefer multiword items (Moon, 1997), formulaic language (Wray, 2002) or formulaic sequences (Schmitt, 2004). Sinclair (1991, 2000) puts phraseology at the heart of language

description. He defines two principles of language organisation: the idiom principle and the open-choice principle. Sinclair claims this as a way of seeing language as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness […]. Virtually all grammars are constructed on the open-choice principle. (1991, pp. 109-110)
Is this arguably a maxim for task development? What choices is the language user aware of? Does she/he know when a unit is complete? Can the language user accurately judge grammaticalness?

In order to account for restraints that are not captured by the open-choice principle, Sinclair proposes the idiom principle: "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (p. 110).

Willis (pp. 144-60) identifies the following four sub-categories of lexical phrases: (1) polywords, which can be learned as if it were a single word, e.g. re-appearing sequences like according to, in my view; (2) frames which are not continuous, less fixed and adaptable: as … as, are not … but; (3) sentences and sentence stems which are often also part of social acts: May I? Do you mind if I …; and (4) patterns, which resemble frames, but only allow for certain predictable range of words in order to be adapted; e.g. relationship collocates with between; relationship can be substituted by other words denoting conflict or resolution: war between nations.

All of types of lexical phrases appear frequently in natural language. Indeed, Hoey (2005) sees lexical phrases and collocations, in particular, as contributing considerably to the naturalness of language (p. 2) and Skehan explains how "communication in real time" is facilitated by using "ready-made elements and chunks" (1992, p. 186). This facilitating function might help to explain why the lexical phrase has been found to be ubiquitous (Erman & Warren, 2000), and much more of a crucial aspect of communicative competence than the knowledge of syntactic rules (Widdowson, 1989, p. 135). All this builds the argument for an area of linguistic knowledge that will be encountered frequently by the learner and required frequently by the learner and hence should be included in any performance statements. Further, we argue if these chunks are stored and retrieved as wholes (Wray, 2002) then descriptors in writing rating scales can be adapted to follow a lexico-grammatical approach and items can be constructed to cue and trigger this area of linguistic knowledge.

References

Alderson, J. C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. London, New York: Continuum.

Cobb, T. (n.d.). Why and how to use frequency lists to learn words. Retrieved from http://www.lextutor.ca/research/

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe. (2009). Relating language examinations to the Common European Framework of Reference for Languages (CEFR): A manual. Strasbourg: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. Text, 20(1), 29-62.

Halliday, M. (1994). An Introduction to functional grammar (2nd ed.). London: Edward Arnold.

Hoey, M. (2005). Lexical priming: A new theory of words and language. London: Routledge.

Laufer, B. & Nation, I. S. P. (1995). Lexical richness in L2 written production: Can it be measured? Applied Linguistics, 16(3), 307-322.

Larsen-Freeman, D. & DeCarrico, J. (2010). Grammar. In N. Schmitt (Ed.), An introduction to applied linguistics (2nd ed., pp. 18-33). London: Hodder Education.

Moon, R. (1997). Fixed expressions and idioms in English: A corpus-based approach. Oxford: Oxford University Press.

Nation, I.S.P. & Beglar, D. (2007). A vocabulary size test. The Language Teacher, 31(7), 9-13.

Sarig, G. (1989). Testing meaning construction: can we do it fairly. Language Testing, 6(1), 77-94.

Schmitt, N. (Ed.). (2004). Formulaic sequences: acquisition, processing and use. Amsterdam/Philadelphia: John Benjamins.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford: Oxford University Press.

Sinclair, J. (2000). Lexical grammar. Naujoji Metodologija, 24, 191-203.

Skehan, P. (1992). Strategies in second language acquisition. In: Thames Valley University Working Papers in English Language Teaching, No. 1.

Widdowson, H. G. (1989). Knowledge of language and ability for use. Applied Linguistis, 10, 128-37.

Willis, D. (2003). Rules, patterns and words: Grammar and lexis in English language teaching. Cambridge: Cambridge University Press.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.

Maria Stathopoulou

University of Athens, Athens, Greece

mastathop@enl.uoa.gr

Investigating Mediation as Translanguaging Practice in a Testing Context: Towards the Development of Levelled Mediation Descriptors

Bio data

Maria Stathopoulou is a graduate of the Faculty of English Language and Literature, University of Athens (honors degree). She holds a PhD (with distinction) and an MA degree in Applied Linguistics (valedictorian) from the same faculty. She is currently working for the Research Centre for Language Teaching, Testing and Assessment, University of Athens, as a research asssociate. Since 2008, she has been a member of the English team preparing the national exams for the Greek State Certificate of Foreign Language Competence (known as KPG). She is also involved in the "Teaching of English to Very Young Learners" project, known as PEAP, and is a member of the research team responsible for the linking of the KPG exams to the mainstream educational system. Additionally, she has been involved in the KPG test-taking strategies project and has helped organize some KPG experimental classes at state institutions. She has also been awarded the Sasakawa Young Leaders Fellowship Fund (Sylff) by The Tokyo Foundation. Her work, which primarily focuses on EFL testing and assessment, the evaluation of mediation performance and the investigation of test-taking strategies, has been presented in national and international conferences. Her recent papers are: The linguistic characteristics of written mediation tasks which is to appear in Versita (Polland) and Test-taking strategies in the KPG reading test: Instrument construction & investigation results, which appears in The Journal of Applied Linguistics (JAL).

Abstract

This paper reports on findings of a longitudinal research project exploring the complex nature of interlinguistic mediation -a communicative undertaking which entails purposeful selection of information by the mediator from a source text in one language and relaying this information into another language, with the intention of bridging the communication gap between interlocutors. Although in today's multilingual contexts, it is essential for individuals to have acquired the skills and strategies that will enable them to use two or more languages in a parallel fashion (an ability foreseen by the Common European Framework of Reference for Languages), mediation as translanguaging practice has not received much attention, probably for reasons related to the monolingual paradigm in mainstream language teaching and testing. Given that the CEFR provides no benchmarked illustrative descriptors relevant to mediation, this research has set out to investigate what counts as successful mediation. Specifically, by drawing data from the KPG English Corpus, which comprises collections of written texts (scripts) produced by users of English who have sat for the Greek national standardized foreign language exams (known as KPG) -the only examination system in Europe which assesses test-takers' mediation ability- this study identifies successful mediation strategies in scripts of different proficiency levels from different KPG writing test papers over a period of six (6) years. The paper actually presents an inductively and empirically derived Inventory of successful mediation strategies which may contribute to the creation of standardized measures and clear benchmarks for reliable assessment of mediation competence thus complementing the CEFR.

Short paper

The focus and context of the research
Motivation for the research

The present paper deals with the notion of (intelinguistic) mediation, which involves relaying in one language messages purposefully extracted from a source text in another language, so as to restore communication gaps between interlocutors. It attempts to define mediation on the basis of results derived from a large-scale research project, which investigated what counts as successful mediation in a testing context (cf. Stathopoulou, 2013)[1] and ultimately stresses the importance of developing levelled descriptors relevant to mediation on the basis of empirical evidence. What is discussed herein is actually based on research which has drawn data from the Greek national foreign language exams (known as KPG), which is the only examination system in Europe which has legitimized mediation by assessing test-takers' mediation competence (cf. Dendrinos, 2006). In fact, consistent with the recommendations of the European Commission and the Council of Europe to promote multilingualism, the KPG exams assess written and oral mediation performance from B1 level onwards thus promoting linguistic diversity (rather than one single language, i.e. English).[2]

To set the context, the aforementioned project is related to recent research in promoting multilingualism and more specifically it is placed within a wide context of ongoing research conducted in Europe engaged in setting standards for language learning and assessment (cf. Green, 2010; Krumm, 2007; Alderson et al, 2004). In this context, scholarship addresses questions such as, What does multilingual literacy (or multilingual competence) entail and on the basis of what criteria can it be assessed (cf. Garcia, Flores and Woodley, 2012; Dendrinos, 2012; Shohamy, 2011; Lenz and Berthele, 2010; Coste and Simon, 2009; García, Barlett and Kleifgen, 2006)? What skills should language learners develop in order to participate effectively in today's multilingual and superdiverse[3] societies (cf. Hornberger and Link, 2012, Hornberger, 2007) and through what foreign language education pedagogies can the ability to use translanguaging4 and interlinguistic mediation techniques be developed (cf. García, Flores and Woodley, 2012; Gort and Pontier, 2012; Hambye and Richards, 2012; Yagmur and Extra, 2011, Creese and Blackledge, 2010)?

What triggered discussions in relation to the aforementioned issues in the field of bilingual education and foreign language pedagogies is the urgent need for communication in the new multilingual environments which impose new realities, challenges and demands on language users. As a consequence, in the new multilingual contexts of social, political and economic struggles (García, 2008: 388) and cultural diversity, people use translanguaging (or polylanguaging[5]) techniques drawing upon the resources they have from a variety of contexts and languages, and ultimately resort to the use of mediation. As a matter of fact, it seems very likely for a person to act as a mediator, i.e., to find himself/herself in a situation in which s/he has to serve as a linguistic and cultural bridge between individuals who do not share the same language

[1] Doctoral research under the supervision of Professor B. Dendrinos, University of Athens. Note that this research is related to the work which is being carried out at the Research Centre for Language Teaching, Testing and Assessment (RCeL) (http://rcel.enl.uoa.gr). This work is co-funded by the European Social Fund and the Greek National State – (NSRF), under the project of the National and Kapodistrian University of Athens entitled "Differentiated and Graded National Foreign Language Exams".

2 For further information concerning the rationale and the underlying ideology of the KPG exams, see Dendrinos (2009). Also visit: http://rcel.enl.uoa.gr/kpg

[3] Within the framework of ethnic, migration, racial and sociology studies, Vertovec (2007, 2009) uses the term of 'superdiversity' to refer to the example of England and particularly London which is "the predominant locus of immigration and it is where super-diversity is at its most marked" (Vertovec, 2007: 1042).

[4] Translanguaging describes the use of literacy practices to "move back and forth with ease and comfort between and among different languages and dialects, different social classes, and different cultural and artistic forms" (Guerra, 2004: 8).

[5] Polylanguaging refers to the use of different linguistic resources associated with different languages available in the user's repertoire (Jørgensen and Møller, 2012; Jørgensen, et al, 2011, Jørgensen, 2010, 2008).

and relay messages from one language to the other for a given communicative goal. Interlingual mediation thus seems to be an important aspect of human intercultural communication that deserves particular attention in any discussion for foreign language testing and appropriate language pedagogies.

The notion of mediation in foreign language didactics became widely known with its inclusion in the Common European Framework of Reference for Languages (henceforth CEFR) (Council of Europe, 2001) which considers mediation activity as an important part of someone's language proficiency. However, it has not received as much attention as the activities of reception, production and interaction. As a matter of fact, no benchmarked illustrative scales for the mediatory use of language are available therein (cf. Alderson 2007, Little 2007, North 2007). Given this void, the particular language activity has seldom been included in foreign language curricula or featured in classroom activities until recently,[6] and its investigation is at embryonic stages.[7]

What thus motivated the research, extensions of which are presented herein, is the need to further explore interlingual mediation, which has been absent in the scene of foreign language testing and teaching probably for reasons related to the monolingual paradigm in mainstream language teaching and testing, which is still real in our days, as Dendrinos (2012) maintains.

Interlingual mediation as translanguaging practice: theoretical considerations

Interlingual mediation is considered as a form of translanguaging as it is a language practice which involves, as Garcia et al. (2011) would put it, a 'hybrid practice of languaging'. Translanguaging, which is also referred to in the literature as 'transcultural repositioning' (Richardson-Bruna, 2007: 235),[8] is a term introduced by Williams (1994, 1996) and refers to the alternation of languages in multiple modes, i.e., spoken and written, receptive and productive (cf. García, 2009a; Baker, 2001a, 2001b; Williams 1994). In 'translanguaging', the input (reading or listening) tends to be in one language and the output (speaking or writing) in the other language. The issue of translanguaging has become commonplace in discussions among scholars dealing with communication within a context of social, political, and economic struggles (García, 2008: 388) unavoidably occurring in today's contexts of linguistic and cultural pluralism. García (2009b) argues that rather than focusing on the language itself, translanguaging makes it clear that there are no clear-cut boundaries between the languages employed. In much the same vein, Canagarajah (2011) points out that multilingual competence emerges out of local practices where multiple languages are negotiated for communication; [...] competence does not consist of separate competencies for each language, but a multicompetence that functions symbiotically for the different languages in language user's repertoire.

In this paper, mediation as translanguaging practice is sharply distinguished from the meaning it takes in the CEFR, which sees it as somehow synonymous with (professional) translation and interpretation (Council of Europe, 2001). Translation requires unconditional respect of the content of the source text, and the aim of the translator or the interpreter is to render every single message of the original text (Dendrinos and Stathopoulou, 2010, 2011). Equally important is the requisite that the target text be in the same textual form as the source text. On the contrary, the aim of the mediator, unlike the translator (or the interpreter), is to select from the source text information relevant to the task at hand and to render it appropriately for the context of situation. In

[6] In Greece, the newly developed National Curriculum for Foreign Languages actually includes illustrative descriptors for the mediatory use of language, which are empirically developed and are partly based on the task-analysis results presented in Stathopoulou (2013) (cf. Dendrinos and Stathopoulou, 2011).

[7] Another research also focusing on the KPG exams has been conducted by Stathopoulou (2009) within the framework of her MA studies at the University of Athens.

[8] 'Transcultural repositioning' describes the use of literacy practices to "move back and forth with ease and comfort between and among different languages and dialects, different social classes, and different cultural and artistic forms" (Guerra, 2004: 8).

other words, while reproduction of a text establishing equivalents between two texts is the very essence of translation, mediation involves relaying of certain pieces of information from a source text to a target text.

Overall, the mediator is viewed as a plurilingual social actor actively participating in the intercultural communicative event, drawing on source language content and shaping new meanings in the target language.

Mediation competence and performance: Towards developing levelled descriptors

In response to the need for further investigation as to what ensures the success of mediation, the research project, several implications of which are discussed in the present paper, has attempted to constitute a step towards shedding light on aspects of this unexplored area. While the aim of the research was to acquire a general understanding of the mechanisms of interlinguistic mediation in a testing context by analysing mediation tasks and texts (i.e., scripts as result of mediation tasks),[9] the aim of this paper is to raise awareness of the gap in research as to what mediation is and to suggest a framework for the development of mediation-specific can-do statements which will include a lexicogrammatical description of mediators' language production. The quantitative and qualitative analysis of mediation tasks and learner corpora (KPG mediation scripts) for the purposes of the aforementioned research project, has led to the development of an empirically and inductively derived framework (the so-called, Inventory of Written Mediation Strategies (IWMS)) (Stathopoulou, 2013), which can be used in the future for the construction of levelled mediation strategy descriptors.

By exploring what successful (written) mediation is through textual analysis, the study, the extensions of which are herein discussed, constitutes a systematic attempt to complement the CEFR by developing objective criteria so as to describe levelled language proficiency, which will in turn facilitate the development of standards in language teaching and testing (cf. Green, 2010; Krumm, 2007; Alderson et al, 2004), intended to help the mutual recognition of qualifications gained in different learning contexts. As a matter of fact, the findings derived from this investigation may contribute to the development of empirically validated descriptors related to the simultaneous use of more than one language.

Any attempt to create mediation specific descriptors could take into account that language users' ability to mediate and translanguage does not only involve being competent in two languages making use of their linguistic knowledge but it also entails being competent in shuttling between languages and in crossing linguistic borders in order to communicate by relaying information from one language to the other according to the rules and possibilities of the communicative encounter (Stathopoulou, 2013). This sort of competence is related to the ability to use a number of different mediation strategies (see examples in Table 1), which are defined as those strategies needed in order to successfully relay information from one language to another for a given communicative purpose.

[9] The corpus included texts having been produced over a period of four (4) years by KPG candidates sitting for the B1, B2 and C1 level exam and the total number of words comprising it was almost 53.000. The RCeL has been digitalizing KPG candidates' scripts since 2004 with a view to developing a corpus which will be used for the investigation of the Greek Foreign Language Learner's Profile (Gotsoulia and Dendrinos, 2011). The corpus now consists of about five million words. A range of A1-C1 level scripts rated as fully satisfactory, moderately satisfactory and unsatisfactory comprise the corpus.

	Mediation Strategies
01.	Creative blending between extracted and extra-textual information
02.	Combining information
03.	Summarising
04.	Reorganising extracted information
05.	Condensing (at sentence level) by combining two (or more) short sentences into one (sentence fusion)
06.	Expanding
07.	Paraphrasing

Table 1: Mediation strategies as presented and defined by Stathopoulou (2013)

To elaborate on the table above, the mediator may combine information from different sources, i.e., his/her background knowledge on a topic (i.e., the case of creative blending between inserted and extracted information) or the source text which is in a different language from the target text (i.e., the case of combining of extracted information). S/he may also reorganise source text sentences or whole paragraphs and may summarize source information to its gist, either through a sentence or through more than one sentence. Additionally, the mediator may use a variety of paraphrasing strategies (i.e., reformulation of the exact words of the source text) both at the level of text and sentence and may expand or condense the initially used sentences. Of course, as research has indicated the aforementioned strategies are not independent of the task. Being thus able to mediate also implies "dealing with task requirements in such a way that the outcome will include -apart from the appropriate language- those mediation strategies conducive to the task at hand, consequently contributing to the success of mediation" (Stathopoulou 2013: 311).

Given thus the inextricable link between task and performance, mediation-specific can-do statements are also important to take into account both task requirements and actual performance. As a matter of fact, any effort undertaken up to now towards the development of mediation-specific descriptors, i.e., the Profile Deutsch,[10] has not taken into consideration the tasks and their demands thus providing descriptors which are not articulated as task-dependent communicative production.

In addition to the above, the mediation-specific descriptors based on empirical evidence should not only specify the mediation strategies needed for learners of different levels when being involved in different mediation tasks but also describe the language to be used by learners at each proficiency level. As a matter of fact, the linguistic documentation of the mediation competence across the CEFR language proficiency levels by systematically analysing the language found in texts produced by mediators of different levels will contribute to the creation of language-specific descriptors, which will add grammatical and lexical details of the target language to CEFR's functional characterization of the different levels (Hawkins and Filipović, 2012: 5).

Conclusion

Empirically validated descriptors for different levels of language proficiency are definitely in demand in order to supplement the rather vague CEFR descriptors or the language proficiency descriptors of various language testing systems and curricula. But descriptors related to the simultaneous use of more than one language, whether in a real-life communicative encounter or in a testing situation, are missing altogether –even in CEFR terms– while studies and research regarding mediation and other multilingual practices

[10] The *Profile Deutsh* (Glaboniat et al, 2005) includes can-do objectives at different proficiency levels, which were set out for the various categories of activity according to their treatment in the CEFR: reception, production, interaction and mediation.

are generally wanting.[11] It is exactly this void that this research was intended to fill, given the lack of objective criteria to describe mediation skills and strategies in the CEFR. The resulting descriptors relevant to mediation could inform mediation task design for testing (or teaching) purposes in the future and could generally constitute the basis for the development of multilingual curricula, language exam specifications, and foreign language materials.

References

Alderson, C. (2007). The CEFR and the need for more research. The Modern Language Journal, 91(4), 659-663.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2004). The Development of Specifications for Item Development and Classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Final Report of The Dutch CEF Construct Project. Retrieved, July 22, 2012 from http://eprints.lancs.ac.uk/44/1/final_report.pdf

Baker, C. (2001a). Foundations of Bilingual Education and Bilingualism (3rd ed.). Clevedon: Multilingual Matters.

Baker, C. (2001b). Education as a site for language contact. Annual Review of Applied Linguistics, 23, 95–112.

Canagarajah, S. A. (2011). Translanguaging in the classroom: Emerging issues for research and pedagogy. Applied Linguistics Review, 2, 1-28

Coste, D. & Simon, D.-L. (2009). The plurilingual social actor. Language, citizenship and education. International Journal of Multilingualism, 6(2), 168-185. Retrieved August 22, 2012, from http://www.coe.int/t/dg4/linguistic/Source/SourcePublications/CompetencePlurilingue09web_en.pdf

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

Creese, A. & Blackledge, A. (2010). Translanguaging in the Bilingual Classroom: A Pedagogy for Learning and Teaching? The Modern Language Journal, 94, i, 103-115.

Dendrinos, B. (2006). Mediation in Communication, Language Teaching and Testing. Journal of Applied Linguistics, 22, 9-35.

Dendrinos, B. (2009). Rationale and Ideology of the KPG Exams. ELT News. Retrieved August 16, 2012, from: http://rcel.enl.uoa.gr/kpg/kpgcorner_sep2009.htm.

Dendrinos, B. (2012). Multi- and monolingualism in foreign language education in Europe. In G. Stickel & M. Carrier (Ed.), Education in Creating a Multilingual Europe. Frankfurt: Peter Lang. 47-60.

Dendrinos, B. In press (Spring 2013). Social meanings in global-glocal language proficiency exams. In C. Tsagari, S. Papadima-Sophocleous & S. Ioannou-Georgiou (Ed.),

[11] Shohamy, who claims that all assessment policies and practices are based on monolingual constructs not allowing other languages to 'smuggle in' (2011: 1), argues in favour of the adoption of different types of multilingual testing and assessment policies and practices. In addition, Dendrinos (in press, 2013) maintains that locally-controlled testing suites may serve as counter-hegemonic alternatives to the profit-driven global language testing industry.

Language Testing and Assessment around the Globe: Achievements and Experiences. Language Testing and Evaluation series. Peter Lang.

Dendrinos, B. & Stathopoulou, M. (2010). Mediation activities: Cross-Language Communication Performance. ELT News, 249(12). Retrieved August 20, 2012, from http://rcel.enl.uoa.gr/kpg/kpgcorner_may2010.htm

Dendrinos, B. & Stathopoulou, M. (2011). Η διαμεσολάβηση ως σημαντική επικοινωνιακή δραστηριότητα. Οδηγός του Εκπαιδευτικού για το Ενιαίο Πρόγραμμα Σπουδών των Ξένων Γλωσσών. Αθήνα: Παιδαγωγικό Ινστιτούτο, Υπουργείο Παιδείας, Δια Βίου Μάθησης και Θρησκευμάτων. Retrieved August 18, 2012, from http://rcel.enl.uoa.gr/xenesglosses/guide_kef6.htm

García, O. (2008). Multilingual language awareness and teacher education. In J. Cenoz & N. H. Hornberger (Ed.), Encyclopedia of Language and Education (6, 385–400). New York: Springer.

García, O. (2009a). Education, multilingualism and translanguaging in the 21st century. In Mohanty, A., Panda M., Phillipson, R., Skutnabb-Kangas, T. (Ed.), Multilingual Education for Social Justice: Globalising the Local (pp. 128-145). New Delhi: Orient Blackswan (former Orient Longman).

García, O. (2009b). Reimagining Bilingualism in Education for the 21st century. Paper presented at the NALDIC conference 17, University of Reading, 14 November.

García, O., Bartlett, L. & Kleifgen, J. From biliteracy to pluriliteracies. (2006). In P. Auer and L. Wei (Ed.), Handbook of Applied Linguistics on Multilingual Communication Vol. 5: Multilingualism (pp. 207-228). Berlin: Mouton/de Gruyter,.

García, O., Flores, N. & Woodley, H. (2012). Transgressing monolingualism and bilingual dualities: Translanguaging pedagogies. In A. Yiakoumetti (Ed.), Harnessing linguistic variation to improve education (pp. 45-76). Bern: Peter Lang.

García, O., Makar, C., Starcevic, M. & Terry, A. (2011). The translanguaging of Latino kindergartners. In J. Rothman & K. Potowski (Ed.), Bilingual Youth: Spanish in English speaking societies (pp. 33-55). Amsterdam: John Benjamins

Glaboniat, Müller, M., Rusch, M. Schmitz, Wertenschlag P. & Lukas H. Langenscheidt. (2005). Profile Deutsch. Lernzielbestimmungen, Kannbeschreibungen und kommunikative Mittel für die Niveaustufen A1, A2, B1, B2, C1 und C2 des "Gemeinsamen europäischen Referenzrahmens für Sprachen". Berlin, München.

Gort, M. & Pontier, R. W. (2012). Exploring bilingual pedagogies in dual language preschool classrooms. Language and Education, 1-23.

Gotsoulia, V. & Dendrinos, B. (2011). Towards a Corpus-based Approach to Modelling Language Production of Foreign Language Learners in Communicative Contexts. In Proceedings of the 8th Recent Advances in Natural Language Processing conference. Hissar, Bulgaria

Green, A. (2010). Requirements for reference level descriptors for English. English Language Profile 1(1), 1-19.

Guerra, J. C. (2004). Emerging representations, situated literacies, and the practice of transcultural repositioning. In M.H. Kells, V. Balester, & V. Villanueva (Ed.), Latino/a discourses: On language, identity, and literacy in education (pp. 7–23). Portsmouth, NH: Heinemann.

Hambye, P. & Richards, M. (2012). The paradoxical visions of multilingualism in education: the ideological dimension of discourses on multilingualism in Belgium and Canada. International Journal of Multilingualism, 9(2), 165-188.

Hawkins, J. & Filipović, L. (2012). Criterial Features in L2 English. Cambridge: Cambridge University Press.

Hornberger, N. H. (2007). Biliteracy, transnationalism, multimodality, and identity: Trajectories across time and space. Linguistics and Education, 18, 325–334.

Hornberger, N. H. & Link, H. (2012). Translanguaging and transnational literacies in multilingual classrooms: a biliteracy lens. International Journal of Bilingual Education and Bilingualism, 15(3), 261-278.

Jørgensen, J. N. (2008). Polylingual languaging around and among children and adolescents. International Journal of Multilingualism, 5(3), 161-176.

Jørgensen, J. N. (2010). Languaging. Nine years of poly-lingual development of young Turkish-Danish grade school students (Vol. 1-2). Copenhagen: University of Copenhagen.

Jørgensen, J. N., Karrebæk, M. S., Madsen, L. M. & Møller, J. S. (2011). Diversities, 13, 2, 23-37. Retrieved July 15, 2012, from www.unesco.org/shs/diversities/vol13/issue2/art2.

Jørgensen, N. & Møller, J. S. (2012). Aspects of Poly-languaging in Superdiversity. Presentation at the Sociolinguistics Symposium 19. Freie Universität, Berlin.

Krumm, H.-J. (2007) Profiles instead of levels: the CEFR and its (ab)uses in the context of migration. The Modern Language Journal, 91(iv), 667-669.

Lenz, P. & Berthele, R. (2010). Assessment in plurilingual and intercultural education. Document prepared for the Policy Forum The right of learners to quality and equity in education - The role of linguistic and intercultural competences. Geneva, Switzerland. November 2-4. Retrieved 10, November, 2012 from http://www.coe.int/t/dg4/linguistic/Source/Source2010_ForumGeneva/Assessment2010_Lenz_EN.pdf

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. The Modern Language Journal, 91(4), 645-653.

North, B. (2007). The CEFR Illustrative descriptor scales. The Modern Language Journal, 91(4), 656-659.

Richardson-Bruna, K. (2007). Traveling tags: The informal literacies of Mexican newcomers in and out of the classroom. Linguistics and Education, 18, 232-57.

Shohamy, E. (2011). Assessing Multilingual Competencies: Adopting Construct Valid Assessment Policies. The Modern Language Journal, 95(iii), 418–429.

Stathopoulou, M. (2009). Written mediation in the KPG exams: Source text regulation resulting in hybrid formations (MA Dissertation). Faculty of English Studies. University of Athens. Retrieved September, 10, 2012 from http://rcel.enl.uoa.gr/kpg/texts/MA%20thesis_Stathopoulou_mediation.pdf

Stathopoulou, M. (2013). Task dependent interlinguistic mediation performance as translanguaging practice: The use to KPG data for an empirically based study.

(Unpublished PhD thesis). Faculty of English Language and Literature. University of Athens.

Vertovec, S. (2007). Super-diversity and its implications. Ethnic and Racial Studies, 30(6), 1024-1054.

Vertovec, S. (2009). Toward post-multiculturalism? Changing communities, conditions and contexts of diversity. International social science journal LXI, 61(1), 83-95.

Williams, C. (1994). Arfarniad o Ddulliau Dysgu ac Addysgu yng Nghyd-destun Addysg Uwchradd Ddwyieithog (Doctoral thesis). Bangor: University of Wales Bangor.

Williams, C. (1996). Secondary education: teaching in the bilingual situation. In , C. Williams, G. Lewis and C. Baker (Ed.), The Language Policy: Taking stock, 12(2), 193–211.

Yagmur, K. & Extra, G. (2011) Urban multilingualism in Europe: educational responses to increasing diversity. Journal of Pragmatics, 43(5), 1185-1195.

Martine Swennen

Delft University of Technology, Delft, The Netherlands

m.a.swennen@tudelft.nl

From a Low-Stakes Test to a Higher-Stakes Test

Bio data

Martine Swennen has taught English as a foreign language at tertiary level for more than 18 years. With MA degrees in both English and Applied Linguistics her interest in teaching and learning languages has been clear from the start of her career. In Delft she teaches courses in English and Presentation skills to both staff and students. Her other professional interest is CALL, Computer Assisted Language Learning. Martine Swennen is co-author of Lesgeven in het Engels (Teaching in English).

Abstract

In order to put our students in the correct course, whose levels correspond to the CEFR levels, our department uses its own Placement Test (a cloze test with two short essay questions). From experience we know that the results are a fair reflection of the participants' language skills needed for our courses.

This test is a so-called low stakes test, e.g. there is no pass/fail cut-off point. However, English as the language of instruction is taking a more important place than before and we receive an increasing number of questions from students asking about an assessment of their level of English, for instance because they want to enter a Master's degree course where English is the language of communication. This means that the stakes are becoming higher and the question is whether we can use our test to determine if a student's English is good enough to participate in such a programme.

If we want to make higher-stakes claims about our test, we need to ensure that the test is accurate and reliable, irrespective of the test-takers' background. We are now at the beginning of the project.

Short paper

In our English department we have been using cloze tests for more than 10 years. Its purpose is simple: to put the test-takers in the correct English course. We use a test consisting of three or four texts with approximately 100 gaps, complemented by two short essay questions asking about the students' field of study and their English skills. Using MapleTA©, we digitalised our test, thus reducing the workload substantially. All in all, we find that the test suits its purpose well.

However, we are receiving a growing number of requests to use our test to assess students' general English skills that are necessary to participate in English-taught university education. This means that we need to investigate our test more closely to see if it can be used to serve this new purpose. The current project was undertaken to do so.

A preliminary literature study revealed that a large number of issues should be considered when carrying out this kind of project (see for example Alderson (1979)). The value of cloze tests has been debated ever since Taylor (1953) started doing research on

them. Many researchers have added to the discussion by looking at methods of deletion and scoring, item analyses and test purpose, for instance Abraham and Chapelle (1992), Jonz (1990) and O'Toole and King (2011). The time spent on our project has been too short to draw any major conclusions, but it is becoming clear that the following issues require further investigation:

1. Deletion method: we have chosen for the fixed ratio method. With a team of teachers this ensures that different people use the same method, thus eliminating the influence of the test maker. A more in-depth analysis of the literature available needs to confirm (or reject) if this is an acceptable method here.
2. Scoring method: so far we have always adopted the "acceptable answer" method, because we felt that this gives both higher and lower level students a similar opportunity to show their English language skills. Brown (1980) puts it like this: "There is something inherently repugnant about counting an answer wrong, which is actually correct, simply because the author of the original passage did not choose to use that word".(p.316) However, much has since been written on this topic, so further research is needed to see if this is still a correct assumption.
3. Test purpose: a number of studies report on the use of cloze tests to determine the difficulty of texts (Benjamin, 2012), or to assess students' reading comprehension (Gellert & Elbro, 2013). Bachman (1982) mentions, that "cloze tests have been found to be highly correlated […] with tests of nearly every language skill and component" (p.61), and Tremblay's research (2011) seems to support this claim. Jonz's (op.cit.) research also indicates that cloze scores correlate with language comprehension. However, for our purpose, it is necessary to see if test results can be indicative for productive skills too.

Another issue at stake in this project is test reliability: test-takers should have an equal chance of reaching a certain score regardless of their language and cultural background and the version of the test they get. For this we need to analyse the tests we have used and will use, focusing on test-takers' backgrounds and the texts that we have used.

A final matter to be investigated is how scores on the test can be interpreted: our English for Academic Purposes course is offered at four different levels ranging from A2 to C1 on the scales of the Common European Framework (CEFR). We therefore need to determine with objective means if the scores on our test can be transferred to the CEFR.

This project seems to be related to three conference themes:

- **Test purpose**
  At the moment we are testing for Placement purposes only. With this project we are trying to determine if we can also use our test for a different purpose, i.e. whether the scores can predict if a student can successfully participate in a university programme where English is the language of communication. We try to do this by means of the second conference theme:
- **Practicality**
  We link our test to the CEFR through our courses: if our courses do relate to the CEFR levels, may we assume that the scores on our test are also related to the CEFR and by following this course of reasoning….
- **Degree of difficulty of the levels**
  May we thus say that a student who scores a certain score has the required language skills?

In the period to come the topics mentioned above will be further investigated to see if it is indeed possible to use our Placement Test for this new purpose.

References

Abraham, R. G. & Chapelle, C. A. (1992). The meaning of Cloze test scores: An item difficulty perspective. The Modern Language Journal, 76(4), 468-479.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. Tesol Quarterly, 13(2), 219-227.

Bachman, L. F. (1982). The Trait Structure of Cloze Test Scores. Tesol Quarterly, 16(1), 61-70.

Benjamin, R. G. (2012). Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. Educational Psychology Review, 24(1), 1-26.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. The Modern Language Journal, 64(3), 311-317.

Gellert, A. S. & Elbro, C. (2013). Cloze Tests May be Quick, But Are They Dirty? Development and Preliminary Validation of a Cloze Test of Reading Comprehension. Journal of Psychoeducational Assessment, 31(1), 16-28.

Jonz, J. (1990). Another turn in the conversation: what does Cloze measure? Tesol Quarterly, 24(1), 61-83.

O'Toole, J. & King, R. (2011). The deceptive mean: Conceptual scoring of cloze entries differentially advantages more able readers. Language Testing, 28(1), 127-144.

Taylor, W. L. (1953)."Cloze procedure": a new tool for measuring readability. Journalism Quarterly, 30.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. Studies in Second Language Acquisition, 33(3), 339-372.

Claire Tardieu, Monique Reichert & Annick Rivens Mompean

Université Sorbonne Nouvelle, Paris, France
University of Luxembourg, Luxembourg, Luxembourg
University of Lille, Villeneuve d'Ascq, France

claire.tardieu@univ-paris3.fr - monique.reichert@uni.lu - annick.rivens@univ-lille3.fr

The e-CLES Project: How to Make a Scenario-Based Certificate Valid, Reliable and Fair?

Bio data

After working for 18 years in the field of initial and in-service training of language teachers, **Claire Tardieu** is currently Professor of English Didactics at the University of Sorbonne Nouvelle-Paris 3, a position she has held for three years. She lectures on foreign language teaching and learning to undergraduates, postgraduates and teacher students. She is assistant director in the English Department, in charge of the Master of Education. An active member of Prismes (Sesylia) research unit, she is responsible for research in didactics in the field of Second language acquisition and learning.

Since 2002, she has been with the Department of Evaluation, Prospective and Performance (DEPP) at the French Ministry of Education. As French expert, she took part in 3 European projects relating to evaluation and certification:

- 2008-2011: Project coordinator of CEF-ESTIM, European Centre For Modern Languages, (CELV), Austria, http://cefestim.ecml.at/
- 2004-2007: French expert in the EBAFLS project: http://www.cito.com/research_and_development/participation_international_research/ebafls.aspx
- 2003-2004: French Expert in the Dutch CEFR Grid project led by Charles Alderson (Lancaster University) http://www.lancs.ac.uk/fss/projects/grid/
- 1998 - 2000: French University Partner in the Lingua D project : Guide for Tandem Language Learning in Secondary Schools.

Monique Reichert is a psychologist specialized in language assessment, cognitive science issues and empirical methods. From 2002 to 2004 she worked as a research assistant at LIFE Research & Consult, where she was particularly involved in research in the domain of second and foreign language learning. She joined the EMACS Research Unit in 2004, where her main areas of work lie in the development and appraisal of language assessment methods. Since 2006, she is responsible for the development of language tests – in particular French and German reading comprehension and C-tests – used in the context of large scale monitoring projects aiming at evaluating specific aspects of the Luxembourg educational system. She has been actively involved in the data analysis and national report writing for PISA 2006 and 2009, and took part in the following European projects relating to evaluation:

- 2004-2007: Luxembourg expert in the EBAFLS project: http://www.cito.com/research_and_development/participation_international_research/ebafls.aspx
- 2005-2007 : Luxembourg expert for the evaluation part of the project "Pour le multilinguisme: exploiter à l'école la diversité des contextes européens" (Project sponsored by the European Commission Socrates Programme (http://www.uni-giessen.de/rom-didaktik/Multilingualism)

- 2008-2011: Project team member of CEF-ESTIM, European Centre For Modern Languages (CELV), Austria, http://cefestim.ecml.at/
- In 2008, she was an expert consultant of the Institut National des Langues for the development of a teacher training aiming at standardized performance ratings of spoken language. She obtained her PhD in May 2011 on the validity of C-tests.

Annick Rivens Mompean is the director of the Language Center of Lille 3 University, which provides a very convenient field of research, for the different steps defined for the research developed. She is also in charge of the development of language policy at university, focussing on the linguistic competences to be developed and on the validation of these competences through a language certificate. Lille 3 is also the coordinating centre for the Nord Pas de Calais regional CLES pole, and Annick Rivens Mompean belongs to the scientific committee for the CLES national coordination. In her research, Annick Rivens Mompean has adopted a pluridisciplinary approach to deal with the different elements that need to be taken into account for the analysis of language learning in an institutional context, which includes the specific question of language testing.

Abstract

The Higher Education Language Skills Certification (French CLES) is a scenario-based language certification accredited by the French Ministry for Higher Education which is linked to the Common European Framework of Reference for Languages (CEFR). It allows students to obtain a certification to testify their skills in 11 different languages and at three different levels: CLES 1, 2, 3 corresponding respectively to levels B1, B2, C1 of the CEFR. CLES enables five language skills to be assessed: listening, reading, writing, speaking and interacting. As such it belongs to the fourth generation in the history of testing: the action-oriented integrative sociolinguistic approach (cf Spolsky 1981, Reichert, 2011, Tardieu 2013).

This presentation will deal with a submitted Franco-Luxembourgian project aiming at framing the conditions for a valid, reliable and fair CLES. The project will notably involve improving the linking to the CEFR levels: The Dutch CEF Grid[1] as well as the CEF-ESTIM grid[2] and the expertise in standard setting gained through the Socrates EBAFLS project[3] will be used. Key reference materials such as the SurveyLang (Cito, 2011) reports and The Manual (A.L.T.E/Council of Europe, 2011) will be solicited. The expertise of the CLES teams of testers shall be improved through the use of new criterial grids and specific tester trainings. A German and English C-Test (Raatz & Klein-Braley, 1982, Reichert, 2011), as well as additional comprehensive language tests will be used to verify the criterion validity of the CLES. Both reliability and validity of the CLES will be verified before and after the adaptation of the procedures linking the CLES to the CEFR. This will eventually allow the documentation of adequate approaches and methods helping to enhance the validity of language tests such as the CLES with regard to the CEFR.

Short paper

This paper will present some of the main aspects of the e-CLES project – a Franco-Luxembourgian project recently submitted to the ANR (Agence Nationale de la Recherche) in France. After presenting the situation of the Certificat de Compétences en Langues de l'Enseignement Supérieur (CLES) (Language Certificate for Higher Education), we will explain the objectives and the methodology of the project aiming at enhancing the validity, reliability and fairness of the CLES. The part of the project dealing

[1] http://www.lancs.ac.uk/fss/projects/grid/
[2] http://cefestim.ecml.at
[3] http://www.cito.com/research_and_development/participation_international_research/ebafls.aspx

with the development of an online version e-CLES will not be dealt with in the current presentation.

The situation of the CLES

Created by an order dated 22 may 2000 (see B.O.E.N n° 25 of the 29th of June 2000[4]), the CLES was revised by an order dated 28 April 2007 (see B.O.E.N. n° 20 of the 17th of May 2007[5]). After two experimental phases in 2000-2002 and 2003-2005, the CLES was officially recognized in 2007 as an alternative to other better known language tests on the market by the French Secretary for Higher Education. Now it seems that the CLES has gained national significance, by reaching about 60 universities or colleges covering 10 main areas all over France (Aix-Marseille, Bordeaux, Grenoble/Savoie, La Réunion, Lyon, Nancy-Metz, Nord Pas de Calais, Paris/Région parisienne, Rennes/Brest, Strasbourg).

For the sole past year (2011-2012) more than 36 000 students enrolled in 10 languages and 454 sessions.

Positive results:

- The number of CLES candidates in constant increase (Bilan CLES 2011-12) (English: 27886 candidates/33 sessions; German: 1532/56; Spanish: 5415/93);
- The number of CLES sessions in constant increase (Bilan CLES 2011-12, p. 7) (222 in 2008 -272 in 2009 - 395 in 2011);
- New demands for being accredited centers from colleges and the business world;
- New demands for being accredited centers from other countries (Morocco, New Zealand).

Notwithstanding these undeniable achievements, the CLES still fails to get full national and European recognition. Several reasons may explain this:

1. The discrepancy between the task-based approach behind the CLES tests and the teaching practices at University: the students are not well prepared to take a scenario-type test;
2. The high level of failure, especially at Master level. This is partly due to the fact that CLES candidates are almost obliged to take CLES2 (to justify a B2 level for the teaching certificate, for instance) even if their level is rather in keeping with CLES1;
3. Lack of visibility and recognition outside France;
4. Procedures for linking the CLES tests to the CEFR can still be improved;
5. Existing monitoring procedures in designing test materials may still be improved;
6. Existing monitoring procedures in accrediting test materials may still be improved (both a local and a national committee do exist, but the roles have to be clearly defined and exam makers cannot be evaluators at the same time);
7. Existing marking and grading procedures need improving as well as special training of markers and marking supervisors. The grids for written and spoken production need revising;
8. The CLES certificate does not inform about the level reached in the five skills, that is, no information is given to the students about their language proficiency profile. If one fails in one skill, he fails the whole test or if one is sufficiently proficient to succeed at a higher level for one or more skills, this will not be mentioned on the certificate. However, depending on the context in which the target language will be used, specific skills (such as oral language skills) are needed rather than others (e.g., skills related with written language). Future employers and the students themselves possibly want to be informed about their respective strengths and weaknesses. This would also be of interest for students so that they know which skills they still have to improve;

[4] http://www.education.gouv.fr/bo/2000/25/sup.htm
[5] http://www.education.gouv.fr/bo/2007/20/MENS0700723A.htm

The e-CLES project will borrow from comprehensive research through analysing existing procedures for marking and grading performances and feedback from both testers and testees (through background questionnaires, interviews, and surveys). This cross methodology is one of the original aspects of the project and should have a beneficial effect on the overall project.

For the verification of the CLES' validity, the project will also use two sorts of tests which do not belong to the same epistemological background: the C-test and the action-oriented scenario type. This will contribute to bridging the gap between two opposite types of testing and to showing their complementarity.

Objectives and methodology of the project

Since the CLES is a language test that is used for making inferences about students' language proficiency, which, in turn, is in general of significant importance for the students' future life, the importance of the reliability and validity, but also of the fairness of the CLES cannot be overestimated. Proving a test's reliability – which refers to the consistency of the measurement (cf. Bachman, 2007; Davies et al., 1999), - means showing the consistency of the scores from one test set (e.g., part of the items from the test) to the scores of another test set from the same, or another test. In the current context, this would mean that the results of the CLES should not vary significantly as a function of, for instance, the test administrator, the marker, or the test session. Test validity – which refers to the meaningfulness and appropriateness of the interpretations that are made on the basis of the results is closely linked to reliability: A test that is not reliable cannot claim to be valid. In the current project, validity of the CLES means that it measures what is described and elaborated in the CEFR and the CLES test guides. Since the question of reaching or not a given CEFR level is among the most crucial ones for the CLES candidates, particularly the decision of certifying them a particular language proficiency level imperatively has to be valid as far as possible. Finally, test fairness, which closely relates to the test's validity, tackles the consequences of the testing for the tested persons or groups (Davies et al., 1999). In the current project, it particularly concerns the question of whether the CLES does not yield different results as a function of the cultural background of the students, i.e., if the CLES functions uniformly across different cultural groups.

The project will aim to make the CLES scientifically more valid, reliable and fair by improving the used standard setting methodology as well as the task providers' and examiners' expertise, and by completing the current methodology by additional standard setting procedures.

We will thus tackle the points mentioned above by:

- improving monitoring procedures both for designing test material and marking and grading students' performances;
- improving standard setting procedures. The expertise of the CLES teams responsible for the marking and grading of spoken and written production and interaction exercises shall be improved through tester training: the training sessions will be adapted and expanded to include assessing writing and speaking skills through recorded sample performances and the use of revised or new criterial evaluation grids. For the receptive skills tests, the standard-setting team members will also be provided a summary of empirical data derived from the performance of a group of test takers in order to complement the task to provide difficulty estimates based exclusively on the perceived features of an item (The Manual, 2011). This will require administering items at different levels within the same scenario to the same students, but will help tackling the difficult issue of determining at what point a B1 level becomes a B2 level. The Dutch CEF Grid as well as the CEF-ESTIM grid and the expertise in standard setting gained through the Socrates EBAFLS project, as well as the use of various reference materials

such as the SurveyLang (Cito, 2011) reports and The Manual (A.L.T.E/Council of Europe, 2011) will be used for the planning and the realization of all standard setting sessions irrespective of the target language skill;

- analyzing and minimizing test bias. The CLES items should not function differently for students from France as compared to students from another country. This also implies that the test measures the same construct, independent from the language and cultural background of the students. Language assessments will be carried out both in France and in Luxembourg to investigate this issue;
- investigating the task-based approach used in the CLES. One of the strengths of the CLES, in terms of its (face) validity, is the use of a task-based approach: task-based language assessments are more in line with task-based language learning, and positive 'washback' effects on instruction may also be expected. However, test difficulty also risks to vary significantly as a result of test content (Alderson, 2000; Clapham, 1996), suggesting that also the CLES might yield different results as a function of the subject on which the respective CLES session focuses. In the current project, this effect of the test construction approach on final results will be analyzed further;
- studying the effect of a specific test design in the spoken interaction subtest. A specific experimentation will consist of measuring the performance variations in spoken interaction in two different conditions: with a peer (as in CLES or Cambridge Certificates) or with an evaluator (as in the Goethe or Cervantes certifications). Here we assume that the peer-to-peer condition is more relevant than the examiner-to-student condition for levels B1 and B2. Whereas the A1 or A2 CEFR level descriptors indicate that the language user still needs help in spoken interaction, these restrictions disappear at B1 and B2. However, a difference of level between peers is likely to have an impact on the interaction: to what extent such an impact will be positive or negative for both participants? Would they perform the same in the examiner-to-student condition?

The mentioned research topics require the comparison of the results of different groups of students (e.g., students from Luxembourg as compared to students from France). This requires that the compared groups will be comparable to each other in terms of the students' basic language proficiency. Another type of language test – a German and an English C-Test (Raatz & Klein-Braley, 1982; also see Reichert, 2011), which will be elaborated at the University of Luxembourg and which will be available online – will be used for controlling for the students' general language proficiency (i.e., creating groups of students with an equal general language proficiency level).

- Since the C-test is known to be an excellent measure of the commonalities of measures of the main language skills, it will also be used to investigate on the criterion validity of the CLES, by analyzing the correlations between the CLES (subtests and global judgment) on the one hand, and the C-test on the other hand. This presupposes that the same students take the CLES and the C-test. In addition, the same C-test will be administered together with other, well-established language proficiency tests to further corroborate the validity of both the CLES and the C-test.
- increasing the success rate through three specific types of improvements:
- The first one consists of offering the possibility of taking a placement test which would be particularly relevant for e-CLES (in the form of the Dialang placement test or, else, the C-Test in case its validity is confirmed in the project). This would enable students to take CLES 1 or CLES 2 according to their capacity and thus reduce the failure rate;
- The second one will focus on developing CLES papers/items of varying difficulty (i.e. targeting different CLES levels) within the same scenario. This would allow the deliverance of a modular certification according to the skills;

- A third experimentation will deal with the question of whether one should keep to the current functioning of the CLES (no compensation between the 5 skills, that is, a fail in one skill means a fail in the whole CLES) or under what conditions compensation would be acceptable. The analyses of the strength of the relation between the CLES on the one hand and the C-test on the other hand – in particular the question whether the C-test as a measure of a global language proficiency will best reflect the global CLES result – will provide new information regarding this question of compensation. Thus, if a student fails in one skill, should he or she be able to compensate this result with a good level in the C-test (and be certified the target level)?

Deliverables
The objectives cited here will aim to elaborate:

- A guide for CLES test developers, elaborating on the different steps of test development;
- A valid, reliable and fair pilot paper test for CLES 1 and 2 in German and in English. This pilot test and the guide with test development procedures will be experimented with students along with usual CLES test development procedures in order to determine the result in terms of functioning improvement.

We believe that:

- If we manage to give evidence of greater validity, reliability and fairness of the CLES;
- If we manage to improve the success rate;
- then the CLES will definitely be more attractive to students, and universities and colleges and will strive to implement new teaching practices more in keeping with the scenario form of the CLES and European recommendations in general.

Finally, since both reliability and validity of the CLES will be verified before and after the adaptation of the procedures linking the CLES to the CEFR, this will eventually allow the documentation of adequate approaches and methods helping to enhance the validity of language tests such as the CLES with regard to the CEFR.

References

Alderson, C. (2000). Assessing reading. Cambridge, UK: Cambridge University Press.
A.L.T.E. (2011). The Manual for Language Test Development and Examining. Retrieved from Council of Europe website:
http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C.E. Turner & C.Doe (Ed.), Language testing reconsidered (pp. 41-72). Ottawa, CA: University of Ottawa Press.

CITO. (2011). SurveyLang Report. First European Survey on Language competences. Retrieved from European Commission, Education and Training website:
http://ec.europa.eu/languages/eslc/docs/en/final-report-escl_en.pdf

Clapham, C. M. (1996). The development of IELTS: a study of the effect of background knowledge on reading comprehension. Cambridge, UK: Cambridge University Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). Dictionary of language testing (Vol. 7). Cambridge, UK: Cambridge University Press.

Raatz, U. & Klein-Braley, C. (1982). The C-test – a modification of the cloze procedure. In T. Culhane, C. Klein-Braley & D. K. Stevenson (Ed.). Practice and problems in language testing IV (pp. 113-138). Colchester, UK: University of Essex.

Reichert, M. (2011). The validity of the C-Test revisited: findings from a multilingual environment. Unpublished dissertation, University of Luxembourg, Luxembourg.

Rivens Mompean, A. (2011). Articuler des dispositifs innovants pour accompagner vers les certifications en langues. Revue Mélanges CRAPEL 32 "Pratiques d'accompagnement(s) des apprenants en présentiel et à distance", pp. 65-83.

Spolsky, B. (1978). Approaches to language testing. Arlington, VA: Center for Applied Linguistics.

Tardieu, C. (2013) Certifier autrement : CLES ou C-Test ?, Colloque international ACEDLE, Apprendre les langues autrement, Nantes, 7-9 juin 2012 (accepté par les Cahiers de l'ACEDLE, parution 2013).

Jennifer Thewissen

University of Louvain, Louvain-la-Neuve, Belgium

jennifer.thewissen@uclouvain.be

The Criterial Power of Accuracy: a Learner Corpus Approach

Bio data

Jennifer Thewissen holds a PhD in English linguistics which she carried out at the Centre for English Corpus Linguistics at the University of Louvain. Her main areas of interest include the error annotation of learner corpora, SLA developmental patterns, the constructs of accuracy and complexity, and how the developmental study of learner corpora can contribute to fleshing out the Common European Framework of Reference for Languages.

Abstract

Although its value is undeniable, the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) has been criticised on a number of issues, some fair (e.g. its strong reliance on teacher intuition), some rather unfair (e.g. the underspecificity of its descriptors, which was deliberate and unavoidable). These caveats have been argued to constitute rather "shaky ground" (Husltijn, 2007) for such a major educational document.

To address these issues, Husltijn (2007: 665-666) rightly claims that there is an "urgent need to test empirically the implications of the CEFR using real L2 learners rather than teachers". We take heed of this need and suggest a learner corpus approach to further specifying the CEFR descriptors for linguistic competence (grammatical accuracy, vocabulary control, orthographic control and punctuation) for learners of L2 English.

The learner data used in this study consist of 223 argumentative learner scripts taken from the International Corpus of Learner English (Granger et al., 2009) (c. 150,000 tokens in total). Each text was submitted to a rigorous rating procedure and assigned to a specific CEFR score, ranging from B1 to C2. Simultaneously to the rating procedure, each script was annotated for errors, i.e. error tagged, following a 40-plus error taxonomy (Dagneaux et al., 2008). Having a CEFR score and an error profile per text meant that it was possible to capture the developmental path followed by each error type in terms of progress, stabilisation and regression (Thewissen, 2013). Carrying out this developmental learner corpus analysis has shed light on a number of error types which constitute potential "criterial features" (Hawkins & Filipović, 2012) for a given CEFR level. Additionally, our analysis of accuracy development raises a number of questions concerning the actual validity of a six-level proficiency scale as proposed by the CEFR.

Short paper

The criterial power of accuracy: a learner corpus approach
Current research practice

The present research project centres around the novel insights that can be gained from adopting a developmental approach to the study of learner language. The value of such an approach is further boosted by the use of learner corpus data, viz. electronic collections of learner data that have been assembled according to specific design criteria (Granger, 2002) and which constitute authentic rather than invented learner production.

My research aimed at analysing written learner corpus data from a developmental perspective so as to identify areas of significant progress, stabilisation and regression. More specifically, I used 223 assignments included in the International Corpus of Learner English (Granger et al., 2009), each of which I annotated for errors according to the Louvain error tagging taxonomy which distinguishes between more than 40 error types (Dagneaux et al., 2008). In addition to being tagged for errors, each learner script was also rated according to the Common European Framework (CEFR) levels and was subsequently assigned a B1, B2, C1 or C2 overall writing score. Having each text tagged for errors as well as rated for proficiency level meant that it was possible to study the development of the construct of accuracy across the lower intermediate (B1) to mastery (C2) proficiency range. In other words, this analysis yielded insights into the development of 40-plus error types such as spelling, tense errors, uncountable noun errors, punctuation errors or lexical errors and the developmental pattern that they exhibit from B1 to B2, B2 to C1 and C1 to C2. For instance, it was found that while learners significantly progressed in their use of spelling between B1 and B2, the other proficiency levels (B2 and C1, C1 and C2) rather showed a stabilisation trend, i.e. no further significant change concerning the number of spelling errors committed.

The results yielded by this learner corpus developmental research have significant implications for the fields of second language acquisition and language testing. So far, second language acquisition research has mainly focused on cross-sectional studies of learner language, i.e. the study of a specific aspect of language at a single point in time, with the inevitable consequence that little is still known about the developmental paths followed in the acquisition of an L2 (Ortega & Byrnes, 2010). In terms of language testing, the study of accuracy across CEFR proficiency levels has enabled me to compare the developmental paths suggested by the CEFR with the actual developmental results that were produced by the learner corpus data. The development of grammar, vocabulary and orthography as it appeared in the 223 scripts was thus compared with the L2 development that is currently presented in the CEFR descriptors for linguistic competence.

To what extent can or should the levels be made more explicit in terms of required vocabulary and grammar?

The current research paper is linked to the "competence and performance" aspect of the LT&CEFR conference and looks more specifically at whether, and if so how, the CEFR levels could be made more explicit in terms of required vocabulary and grammar.

Because it was mainly developed intuitively, the CEFR document, along with its descriptors, has been found to display a number of oddities and inconsistencies. To give just two examples, my in-depth analysis of the descriptors for grammar, vocabulary, and orthographic control showed that while grammar distinguishes between finer levels B1 and B1+ as well as B2 and B2+, vocabulary and orthography, for their part, choose not to do so and prefer to describe performance at the macro B1 and B2 levels. Another noteworthy point was that, despite its official promotion of the 'can do' approach, the CEFR descriptors are often couched in implicit, and even sometimes explicit, 'cannot do' phrasings. Such preliminary general observations show that the descriptors for grammar and vocabulary do indeed need fine-tuning.

In terms of the developmental paths mentioned in the CEFR and those revealed by the learner corpus data, a number of major differences were noted. For example, the orthographic descriptors for spelling errors are currently worded as follows in the CEFR:

Spelling: accurate enough (B1)→ reasonably accurate but with L1 influence (B2)→ accurate but slips of the pen (C1)→ orthographically free of error (C2)

In terms of development, the CEFR implies spelling to lack any visible development between B1 and B2 where orthographic control is each time qualified as more or less

accurate (i.e. "accurate enough" at B1 and "reasonably accurate" at B2, which intrinsically means the same thing). Progress seems to be located at the later levels, with C1 including a few slips of the pen and C2 being orthographically free of error. The learner corpus-derived developmental results for orthographic development in L2 English led to a different description, however. The learner corpus data actually located the marked progress area to be between B1 and B2, while the stabilisation tendency occurred between B2 and C2, not B1 and B2 as is currently suggested in the descriptors. The difference in developmental patterns is presented graphically in Figure 1 below:

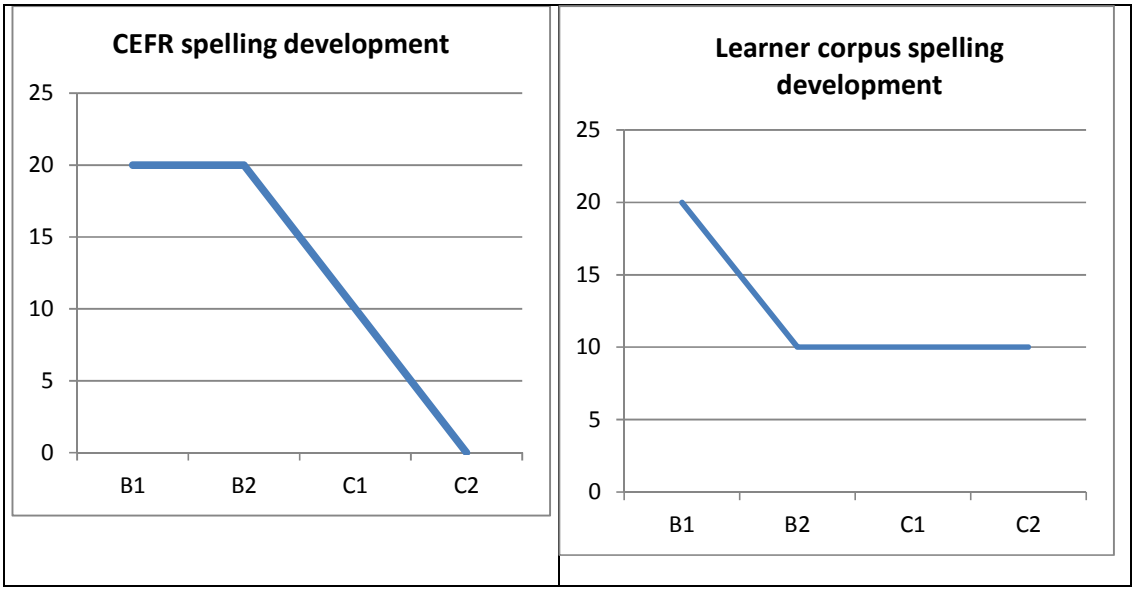


Figure 1: The development of spelling errors: the CEFR vs. ICLE learner corpus data

Results such as these show that relying on teacher intuition to describe L2 developmental paths is an overall unreliable procedure that is likely not to tally with developmental reality. The field of linguistics is fortunate enough to have a considerable number of learner corpora at its disposal and the present study has hopefully shown that language testing research is worth pursuing on this basis.

References

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. & Thewissen, J. (2008). Error Tagging Manual Version 1.3. Centre for English Corpus Linguistics. Université catholique de Louvain, Louvain-la-Neuve. Unpublished manual.

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Ed.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching (pp. 3-33). Amsterdam and Philadelphia: Benjamins.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009). The International Corpus of Learner English. Handbook and CD-ROM (second edition). Louvain-la-Neuve: Presses Universitaires de Louvain. Available from http://www.i6doc.com

Hawkins, J. & Filipović, L. (2012). Criterial Features in the L2 English: Specifying the Reference Levels of the Common European Framework (English Profile Studies). Cambridge: Cambridge University Press.

Hulstijn, J. (2007). The shaky ground beneath the CEFR: Quantitative and Qualitative dimensions of language proficiency. The Modern Language Journal,91, 663-667.

Ortega, L. & Byrnes, H. (2010). The longitudinal study of advanced L2 capacities: An introduction. In L. Ortega, & H. Byrnes (Ed.), The Longitudinal Study of Advanced L2 Capacities (pp. 3-19). New York/ London: Routledge.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error tagged EFL learner corpus. In Modern Language Journal (special issue on capturing the dynamics of L2 development through learner corpus analysis).

Ülle Türk & Tõnu Tender

University of Tartu, Tartu, Estonia

ylle.tyrk@mil.ee - tonu.tender@hm.ee

Re-Designing the School-Leaving Foreign Language Examinations in Estonia

Bio data

Ülle Türk is currently the head of the Language Testing Section of the Estonian Defence Forces' Language Centre at the Estonian National Defence College while continuing to work as a lecturer and teacher educator at the Department of English, University of Tartu. She has been involved in the development of the Estonian school-leaving examinations in English and has been training both teachers and examiners in best practices in language assessment. Her main research interest is language assessment, particularly the testing of reading and writing skills and the issues of standard setting, scaling and calibration.

Tõnu Tender works as an advisor at the Language Policy Department of the Estonian Ministry of Education and Research. His PhD dealt with the issue of multilingualism in the Estonian context. His research areas include language policy, multilingualism and plurilingualism, motivation for language learning and Estonian slang (prisoners' slang, soldiers' slang and students' slang).

Abstract

The Estonian National Curriculum stipulates that all students have to study at least two foreign languages. The choice is usually made from among the four most popular foreign languages: English, Russian, German and French. Until 2014, the foreign language examination was not compulsory though it was the most popular choice with most of the school-leavers opting for English. The examinations developed by the National Examination and Qualification Centre were supposed to be at level B2 roughly. The introduction of a new National Curriculum in 2012 has led to three important changes: 1) from 2014 onwards a school-leaving examination in a foreign language will be one of the three compulsory school-leaving examinations (the other two being the Estonian language and maths); 2) as students can study foreign languages at either B1 or B2 level, they must have a choice as to which of the levels they want to be tested on; 3) the passmark for the national examinations will be abolished – all school-leavers must take the three compulsory examinations, but they can graduate irrespective of the number of points they gain. These changes have meant that the whole examination system will be overhauled. In the case of foreign languages, two separate examination systems will be introduced. No local examinations in Russian, German and French will be developed and students will be expected to take international examinations. However, as the number of students taking English is large, there will still be a locally designed examination in English. A decision has been made to introduce a bi-level examination at levels B1 and B2. The presentation will discuss the reasoning behind these decisions and will look at the measures to be taken to ensure that the locally designed English examination will indeed be measuring students' English language competence at the designated levels.

Short paper

Since 1997, Estonian school-leavers have had to take three national, centrally developed, administered and marked school-leaving examinations of which the only compulsory one has been that of the Estonian language. The other choices have depended on students' preferences, which in their turn have usually been influenced by the entry requirements of the university departments they plan to study at. As most university departments have had proficiency in a foreign language as one of their entry requirements, school-leaving examinations in foreign languages, particularly in English, have been the most popular choices. However, all that is going to change in 2014 as from then onwards there will be three compulsory school-leaving examinations – those in Estonian, mathematics and a foreign language.

The situation with the foreign language examinations is complicated as the National Curricula for the Basic School and for the Gymnasium stipulate that all students have to study two foreign languages. The first of them (FL A) must be taken up in the first school stage (usually in Year 3) and the second (FL B) in the second school stage (usually in Year 6). The most common languages studied are English (as FL A), Russian (as FL B), German and French, though all in all 16 different languages are being taught. By the end of Basic School, students are expected to reach level B1 in FL A and level A2 in FL B. At the Gymnasium (the last three years of secondary education), consequently, they can choose to study these languages at either level B1 or level B2. Thus the school-leaving examinations should allow students to demonstrate their language skills at either of these levels.

Until now, the examinations developed by the National Examination and Qualification Centre are supposed to have been at level B2 roughly. However, no research has been carried out to determine whether this is really the case. Anecdotal evidence suggests that, at least in the case of the English examination, the language competence of those students who have gained at least 80-85 points out of the possible 100 (about 25-30 per cent of the test-taking population) is at level C1. This estimate seems to be consistent with the results of the first European Survey on Language Competences (SurveyLang) according to which in Year 9 (the final year of Basic School) more than 40 per cent of the Estonian test-takers were at level B2 in reading and listening and about 30% in writing (European Commission, 2012, p. 212-213).

Thus there is a question at which level should the school-leaving examination be. The answer to the question lies in the purposes the examination results are supposed to serve. Will they be used just for checking that the curricular aims have been met or should they also indicate how good the foreign language skills of Estonian school-leavers are?

However, due to the smallness of Estonia, it has been deemed impractical to develop two examinations even in the four most commonly taught and studied foreign languages. Thus the Ministry of Education and Research has decided to use internationally recognized examination certificates for school-leaving purposes. From 2014 onwards, no examinations in Russian, German and French will be developed in Estonia. Instead, in the case of Russian, the certificates in Тест по русскому языку как иностранному (B1) and Тест по русскому языку как иностранному (B2) will be recognized. In the case of German, the certificates recognized will include Goethe-Zertifikat (B1), Goethe-Zertifikat (B2), Goethe-Zertifikat (C1), Goethe-Zertifikat (C2) and TestDaF (B2 or C1) and in the case of French those in DELF (B1), DELF (B2) and DALF (C1). The Ministry will pay the fees of those students who decide to take one of the above-mentioned examinations.

As the majority of students have chosen to take the English examination so far and there is no reason to believe that it will change in the future, there will still be a locally developed school-leaving examination in English even after 2014. At the same time,

students can choose to take an internationally recognized examination instead, but they will have to cover the costs themselves. The recognized examinations include those students can take in Estonia, namely the Preliminary English Test (PET), the First Certificate in English (FCE), the Certificate in Advanced English (CAE), the Certificate of Proficiency in English (CPE), the International English Language Testing System (IELTS, at least 4.0) and the Test of English as a Foreign Language (TOEFL iBT, at least 57 points).

Those new regulations mean that the school-leaving examination in English will have to be redesigned. First, it should be possible to determine on the basis of the examination results what level the student's competence in English is at. This means that the examination tasks must be linked to the required levels. Second, as the new curricula do not specify the language structures to be taught, the new examination can only have the four skills parts instead of the five parts (listening and reading comprehension, writing, speaking and language structures) it has had so far.

In order, to ensure that the new examination will test students' competence in English at levels B1 and B2 and that it is possible to compare the results of the new and old examinations, a pilot study is being conducted. It consists of the following stages:

1. The specifications for the new examination were developed. In the case of the writing paper, only the rating scale will change. In the case of speaking, both the tasks and the rating scale will have to be modified. In the case of listening and reading, the tasks will have to be developed at the two targeted levels.
2. On the basis of the specifications, new tasks were developed for speaking, listening and reading.
3. The speaking tasks were tried out on a small number of students and the recordings will be used for developing and trialling of the rating scale as well as for the training of the examiners and raters.
4. The listening and reading tasks were developed.
5. Two standard-setting sessions were held to choose from among the developed listening and reading tasks the ones that will be used in the pilot examination.
6. In April, a pilot examination will be held to determine the levels of difficulty of the developed listening and reading tasks, to see how the new format will work in the case of these two skills and how the results will relate to those of the examination administered this year.
7. Once the results of the pilot examination have been analysed, two more standard-setting sessions will be held to determine the level boundaries for the listening and reading comprehension tests.

As this is the first time standard-setting at this level has been attempted in Estonia, there are numerous concerns and challenges. Primary among them is the fact that the foreign language teachers' familiarity with the Common Reference Levels is quite perfunctory. This has two consequences. First, it is not possible to use student-centred standard-setting methods, and, second, the expertise of the teachers used in the standard-setting process is a serious issue. As there is also little understanding of the standard-setting process and its purposes, it is clear that extra effort must be put into the training of the experts and validating the standard-setting results. The procedures recommended in the Manual for relating language examinations to the CEFR (Council of Europe, 2009) have been (and will continue to be) used with the focus on the familiarization, standardisation training and benchmarking, and validation stages.

References

Council of Europe. (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A

Manual. Strasbourg: Language Policy Division. Available from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp#Manual.

European Commission. (2012). First European Survey on Language Competences: Final Report. Retrieved from http://crell.jrc.ec.europa.eu/eslc_dataset/ESLC_Final%20Report_210612.pdf.

Jan Van Maele & Lut Baten

Katholieke Universiteit Leuven-Group T, Leuven, Belgium
Katholieke Universiteit Leuven, Leuven, Belgium

jan.van.maele@groupt.be - lut.baten@ilt.kuleuven.be

Increasing the Applicability of CEFR Descriptor Scales by Bringing the Context Back into the Framework : Practices from the WebCEF and CEFcult Projects

Bio data

Jan Van Maele is a professor at the Faculty of Engineering Technology, KU Leuven (University of Leuven), Campus Group T, where he teaches English communication courses. He has conducted research in the assessment of language proficiency and intercultural communication, collaborating in the E.U. Lifelong Learning Programme projects WebCEF, CEFcult, and IEREST. He holds a MSc in TESOL from Central Connecticut State University (1988) and a PhD in language testing from the Université catholique de Louvain (1999) with a dissertation on the assessment of oral foreign language proficiency. He has extensive experience in language research and education in China and Belgium as well as in educational management as the adviser to the President of GROUP T–International University College Leuven for strategy, internationalization and communication (1999-2008).

Lut Baten is a full professor at the Leuven Language Institute, KU Leuven (University of Leuven), Belgium, where she was responsible for the courses of English until January 2013, teaching Business English Language and Communication. She obtained a PhD degree at the University of Illinois, Urbana-Champaign and also lectured at the Université catholique de Louvain, Belgium, and at the Mercator University, Duisburg, Germany. From 1975 till 2005 she coordinated Teaching Dutch as Foreign Language in the teacher training department at KULeuven, where she also taught ESP. She has widely published and lectured about ESP, CALL and ICC. She has now retired from teaching, fully concentrating on project management and research pertaining to capacity building in Cuba, South Africa and Congo for the Flemish Interuniversity Council and to several European Lifelong Learning projects, including WebCEF, CEFcult, and IEREST.

Abstract

Citing its deliberate context-free stance, the CEFR presents itself as a framework that is useful for all forms of subjective assessment in self-directed learning (Council of Europe, 2001; North, 2008). At the same time the CEFR authors contend that the framework must remain translatable to each and every relevant context. However, the lack of definition in the CEFR descriptors has proven to present serious challenges even to item writers (De Jong & Jones, 2010). In order to safeguard the usefulness of the CEFR for autonomous learning, we must hence find more effective ways of making descriptors readily accessible.

In this paper the authors will introduce and illustrate three practices for increasing the applicability of descriptor scales by bringing the context back into the framework: (1) including illustrative samples that embody distinctive features of different levels of the scales; (2) adding task-specific descriptors alongside the more abstract official ones; and (3) eliciting feedback in the form of annotations that relate assessments to features of the performance proper.

These practices were originally tested in two European Commission-supported projects, WebCEF (2006-2009) and CEFcult (2009-2011), which centered on developing online assessment platforms for oral skills in relation to the CEFR and INCA scales (INCA, 2004; Baten et al, 2013; Van Maele et al, 2013). Illustrations will be provided for two target use domains: the online job screening interview (for business students in Leuven, Belgium) and the oral presentation of doctoral proposals (for researchers in a program of the Flemish Interuniversity Council and a university in Cuba). To demonstrate the transferability of the presented practices, reference will be made to work-in-progress. The authors will conclude by suggesting that bringing the context back into the framework is a way of honoring the original vision of the CEFR as an on-going exercise in social moderation.

Short paper

Introduction

Citing its deliberate context-free stance, the CEFR presents itself as a framework that is open and flexible in order to be useful for all forms of subjective assessment in self-directed learning (Council of Europe, 2001:6-7; North, 2008:41). At the same time the CEFR authors contend that the framework must remain translatable to each and every relevant context. However, the lack of definition in the CEFR descriptors has proven to present serious challenges even to item writers (De Jong & Jones, 2010). In order to safeguard the usefulness of the CEFR for autonomous learning, we must hence find more effective ways of making descriptors readily accessible.

In this short paper we shall introduce three practices for increasing the applicability of descriptor scales by bringing the context back into the framework: (1) the elicitation of feedback in the form of annotations that relate the assessments to features of the performance proper; (2) the inclusion of illustrative samples that embody distinctive features of different levels of the scales; and (3) the addition of task-specific descriptors alongside the more abstract official ones. Consequently, we shall briefly discuss the manner in which these practices may heighten the practicality of the assessment framework for assessors and assessees outside of the language teaching and testing profession.

These three practices were originally tested in two European Commission-supported projects, *WebCEF* (2006-2009; www.webcef.eu) and *CEFcult* (2009-2011; www.cefcult.eu), which centered on developing online assessment platforms for oral skills in the foreign language (Baten, Beaven, Osborne & Van Maele, 2013; Van Maele, Baten, Beaven & Rajagopal, 2013). More specifically, both assessment platforms aim at developing the users' capacity to assess a sample of oral production against the rating scales, be it one's own speech or that of other learners, and to stimulate self-reflection and dialogue about perceived quality in foreign language proficiency or, in the case of CEFcult, intercultural communicative competence. In accordance with Byram (1997), the latter construct comprises intercultural as well as communicative competence. Since the CEFR does not contain any scales for intercultural competence and given that one of its main authors has indicated that the single scale for sociolinguistic competence that is included ('sociolinguistic appropriateness') is flawed (North, 2008), CEFcult makes use of the INCA scales for intercultural competence in an attempt to fill this gap (INCA, 2004; www.incaproject.org).

Practices

Illustrations for the three practices – annotated feedback, illustrative samples, and task-specific descriptors - will be provided for tasks related to two target use domains: the online job screening interview (for business students at the University of Leuven, Belgium) and the oral presentation of doctoral proposals (for researchers in a development cooperation program of the Flemish Interuniversity Council and the University 'Marta Abreu' of Las Villas, Cuba). Whereas the situation is different for both

groups, the parameters of perception are not: students learn to perceive how their virtual interlocutors see them (perspective, time, place) and thus how to improve the image they wish to convey.

The tasks in WebCEF and CEFcult were designed to elicit, respectively, language samples and intercultural behavior through linguistic performance in those two target use domains. The following rating scales were used in connection with the tasks that are referred to in this paper:

- CEFR: the qualitative aspects of spoken language use: range, accuracy, fluency, interaction, coherence (Council of Europe 2001:28-29);
- CEFR: overall scales for spoken production or interaction (Council of Europe 2001:58, 74);
- INCA: communicative awareness (as a scale for sociolinguistic competence); and knowledge discovery, respect for otherness, and tolerance for ambiguity (as scales for intercultural competence) (INCA 2004:9-10).

The first practice for bringing context back into the framework is to provide free-text annotation boxes through which assessors can relate their assessments to specific features of the actual performance. Both WebCEF and CEFcult contain this function, requesting assessors to make explicit the reasoning behind their evaluations. As can be seen in Figure 1, these annotations can even reveal aspects that, although not covered by the rating scales, may well influence the overall assessment of the speaker's language proficiency ("bonne prononciation"). The WebCEF environment provides the additional feature of time stamping so that certain comments can be attached to selected segments of the recording. In this way, the self-directed learner can discover what the components, the levels, and the descriptors of the framework mean in specific language contexts.

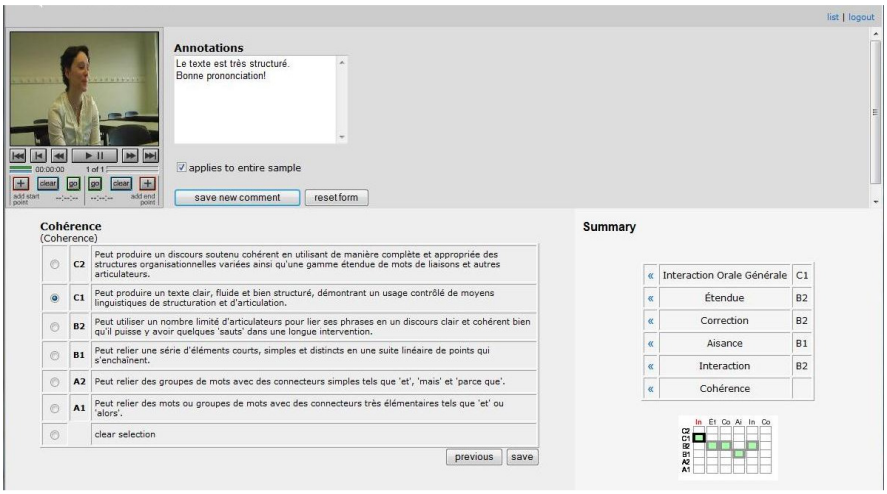


Figure 1. Eliciting annotations in WebCEF

The second practice is to include illustrative samples that embody distinctive features of the different levels in the rating scale. In WebCEF this practice has adopted the form of a showcase, in which each sample represents a different CEFR level of the target language. The samples come with a synthesis assessment, consisting of the consensus ratings assigned by a number of experienced assessors along with the main comments that were cited by these assessors in the annotation boxes. By watching typical performances of a task at various levels and studying the accompanying annotations, learners familiarize themselves with the different scales and levels. In CEFcult this practice is not embedded in a showcase but in situation-oriented observation scenarios. For example, in the online job screening observation scenario, learners assess the intercultural communicative

competence of applicants who are interviewed by someone from a culturally different background. First, learners are instructed to observe the verbal and nonverbal behavior of the applicant non-judgmentally and become aware of what 'is'. Next, they are asked to assess the applicants' performance by matching it with one of the levels in the rating scale concerned. Finally, they are invited to compare their assessment with the interviewer's original, annotated assessment. In this way, the illustrative samples in the WebCEF showcase and the CEFcult observation scenarios present two ways of specifying the levels of the CEFR and INCA scales for all those learners who are not used to linking the abstract descriptors to specific instances of language performance.

The third practice for bringing context back into the framework that we have implemented is to add task-specific descriptors to the official ones. Task-specific descriptors express targeted competence in terms that are more closely related to the test task at hand, as is illustrated in Figure 2 for the INCA scale for 'tolerance of ambiguity'. Given its abstract and opaque phrasing, this scale can be expected to present a serious applicability problem for the novice assessor, as the descriptor for so-called full proficiency demonstrates: "Is constantly aware of the possibility of ambiguity. When it occurs, he/she tolerates and manages it." The accompanying task-specific descriptors, by contrast, render the scale operational by applying the official descriptor to the task concerned, which in this case is to respond to a questionable inquiry after the applicant's attitude towards combining career ambitions with a desire to have children. On top of that, the interviewer's annotated assessment provides additional contextualisation of the descriptor for the learner:

"The applicant does not just cope with ambiguity; she heightens the ambiguity of the situation as a communication strategy. Her opening line – I have the experience that the world is full of surprises – develops into an effective argumentation that allows her to avoid giving a yes or no answer. At the same time she makes clear she does not want to discuss this issue in the context of this interview."

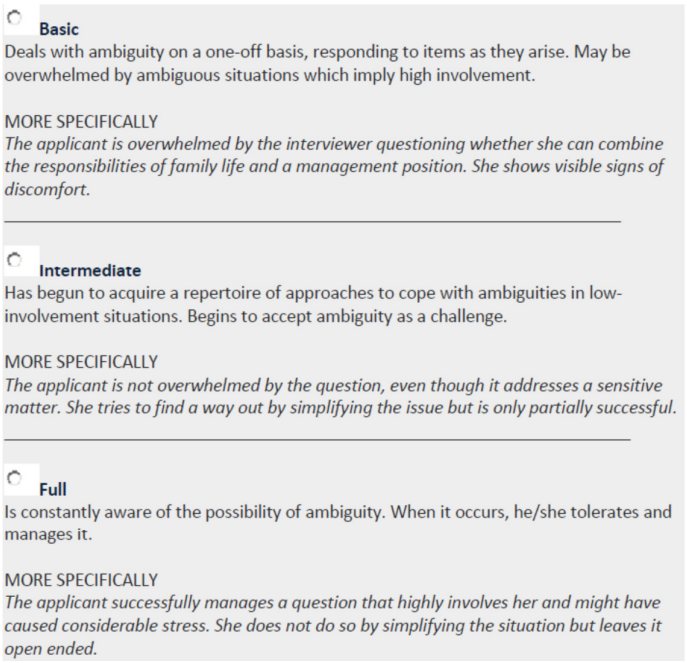


Figure 2. INCA scales and task-specific descriptors for tolerance of ambiguity in CEFcult

Practicality

How applicable is the CEFR to concrete assessment situations? In the context of self-directed learning and collaborative assessment that characterizes the WebCEF and CEFcult environments, this question can only be answered affirmatively insofar as the

CEFR descriptors are useful to the lay assessor, that is nearly everybody outside the language teaching and testing profession. The issue of applicability therefore depends on how successful we as language testers are in bridging the gap between the assessment framework and the novice user. The practices that were introduced above illustrate ways of mediating between the framework and the lay assessor. They present different avenues for increasing the applicability of rating scales by relating the abstract language of the descriptors to specific instances in concrete situations. In this way, these practices demonstrate how features of an online assessment environment can initiate self-directed learners into the assessment framework that they will be operating with.

Practices like the ones that we described do not only mediate by providing stepping stones that lead straight to the framework. They can also, somewhat paradoxically, heighten the applicability of the framework by reaching beyond it. For instance, we observed how the free-text annotations boxes regularly provide an outlet for assessors to comment on features that are not explicitly dealt with in the CEFR scales but might nevertheless impact the assessment, such as accent and pronunciation, intonation, attitude, gestures, and other intercultural aspects of communication style. An optional wizard in WebCEF can channel some of these remarks by offering additional statements to the assessor, e.g.: 'The speaker's body language and appearance enhance the effectiveness of the presentation'.

We also discovered that applicability of descriptor scales is not necessarily a function of the extent to which assessors agree in their judgments. To the contrary, disagreement between raters can be welcomed as a window into how the learner's performance is perceived by a variety of others, particularly in the case of intercultural competence, where appropriateness lies in the eye of the beholder. Assessors may well have valid reasons for disagreeing.

Conclusion

Drawing on our practice as language teachers and testers, we presented three practices in this paper for bringing the context back into the framework, namely annotated feedback, illustrative samples, and task-specific descriptors. In our view, bringing the context back into the framework is not just a way of increasing the applicability of the CEFR descriptor scales, though. It is also a way of honoring the original vision of the CEFR as an on-going exercise in "social moderation" (North, 2010). In spite of the fact that the descriptors were clearly designated as illustrative and were intended as a starting point to encourage investigation, reflection and debate, there has been a distinct and persistent pull towards reification of the scales. The three practices that were illustrated in this paper, by contrast, represent a dialogical and collaborative approach that places assessment-for-learning at the center. By distributing power among all the stakeholders, WebCEF and CEFcult can promote the social process of defining what the CEFR and INCA levels mean. In this manner, we contend, these assessment environments can pave the way for a more "democrative perspective" (Shohamy, 2001) towards assessment that is in keeping with the original view that the CEFR arose from.

References

Baten, L., Beaven, A., Osborne, J. & Van Maele, J. (2013). WebCEF: an online collaboration tool for assessing oral foreign language proficiency. In P.M. Pumilia-Gnarini, E. Favaron, E. Pacetti, J. Bishop & L. Guerra (Ed.), Handbook of Research on Didactic Strategies and Technologies for Education: Incorporating Advancements (pp. 559-570). Hershey, PA: IGI Global.

Byram, M. (1997). Teaching and assessing intercultural communicative competence. Clevedon: Multilingual Matters.

CEFcult. Retrieved March 22, 2013 from http://www.cefcult.eu.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge: Cambridge University Press.

De Jong, J. & Jones, G. (April, 2010). Getting the levels right: deriving item writer guidelines from CEFR descriptors. In 32nd Language Testing Research Colloquium, conducted at the University of Cambridge, Cambridge.

INCA Assessor Manual. (2004). Retrieved March 22, 2013 from http://www.incaproject.org/en_downloads/21_INCA_Assessor_Manual_eng_final.pdf/

North, B. (2008). The CEFR levels and descriptor scales. In L. Taylor & C.J. Weir (Ed.), Multilingualism and assessment. Studies in Language Testing 27 (pp. 21-66). Cambridge: Cambridge University Press.

North, B. (2010). The educational and social impact of the CEFR. In L. Taylor & C.J. Weir (Ed.), Language testing matters: investigating the wider social and educational impact of assessment. Studies in Language Testing 31 (pp. 357-377). Cambridge: Cambridge University Press.

Shohamy, E. (2001). The power of tests. A critical perspective on the uses of language tests. Harlow: Pearson Education.

Van Maele, J., Baten, L., Beaven, A. & Rajagopal, K. (2013). e-Assessment for Learning: Gaining insight in language learning with online assessment environments. In B. Zou, M. Xing, Y. Wang, M. Sun, & C. Xiang (Ed.), Computer-Assisted Foreign Language Teaching and Learning: Technological Advances (pp. 245-261). Hershey, PA: IGI Global.

WebCEF. Retrieved March 22, 2013 from http://www.webcef.eu

Jane Vinther

University of Southern Denmark, Kolding, Denmark

jvinther@sdu.dk

CEFR in a Critical Light

Bio data

Jane Vinther holds an MA in English and Pedagogy and a PhD in computer-assisted language learning. She has extensive teaching and research experience in language and culture, second-language acquisition, cognitive processes of learning, discourse and communication, and philosophy of education. Currently, she holds the position of Head of Department of English Kolding.

Abstract

This presentation takes a critical stance in relation to the CEFR and the 'can do' statements. The principle behind the framework implies that the 'can-do' statements are unitarily understandable and can be interpreted in only one way which will be the same for everyone in every European country. This is probably a fallacy. The political dimension of the CEFR is equally debatable. The convergence towards one system for all disregards national differences and different traditions in educational policy and philosophy. The ecological validity of such a system appears tenuous.

The contemplation of a straightforward set of connectors between a given can-do statement and a corresponding testing framework will lead to a myriad of decisions, all of which will lead to cutting corners and some degree of simplification. The very broad and general statements function at a meta level, and the linguistic expressions or components which will warrant a statement such as "I can with most situations likely to arise whilst travelling in an area where the language is spoken" seem opaque. Furthermore, what does it say about a person's proficiency in relation to the following statement: "I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible". Which of these two statements express the highest level of proficiency? It depends among other things on what the criteria are and how the statements are interpreted. Additionally, the difficulty is to find appropriate linguistic elements to be tested for. None of the statements lend themselves to obvious linguistic measurements.

This paper presents data which will throw light on the difficulty of interpreting the 'can-do' statements at face value, and how difficult even linguists find it to place the statements in a hierarchy.

Short paper

Introduction

Cooperation and common standards are viewed by many as a sign of progress and subsequently as something which makes life easier and better for students, educators and policymakers alike.

One such system is the Common European Framework of Reference for languages. Since its inception in 2001, it has become a household name in a large number of countries, and its use and influence has increased. The Council of Europe, under whose auspices the

CEFR was developed, currently counts 47 member countries , and as this number has expanded so has the dissemination of the CEFR. In 2006 the CEFR was available in 37 languages (Council of Europe, 2007a, 6).

The Council of Europe 2007 Forum Report stated "[…] the clear success of the CEFR has significantly changed the context in which language teaching and assessment of language learning outcomes now take place in Europe." (5). There seems to be grounds for the claims of success not just within a European context but indeed further afield. The diversity of countries in which the CEFR plays a role in one way or another in language educational testing or policy making becomes clear from Byram and Parmenter (2012) in which there are chapter on the influence of the CEFR in countries such as New Zealand, Japan, Taiwan, and the USA, to mention some.

The stated aim of the CEFR is be a reference point for policy and curriculum development and to be a catalyst of increasing cooperation in the educational sphere. In order to ease mobility and transferability of credits across Europe between the educational institutions and programmes, the member states have increasingly incorporated references to the CEFR in official governmental documents (Council of Europe 2007a, 6).

Doubtless, the CEFR is useful and convenient, and appropriately applicable for several purposes. Yet, it is not a universally purposeful tool. The limitations of the policy of unification as well as the CEFR may be overlooked in the desire for convergence and convenience. Especially, the force towards convergence and the streamlined emphasis on skills are concerns that need to be discussed. At the advanced levels in particular the proficiency required should be multifaceted and the worry is that simplified testing rather than comprehensive evaluation should become predominant due to convenience and official promotion of the CEFR. In the following some of those concerns will be voiced and the results of a study into the assessment of the levels will be presented with a discussion of the interpretative nature of the descriptors employed in the CEFR writing frames.

The scope of the CEFR

The declared object of the political body of the Committee of Ministers has been to make the CEFR a vehicle for "strategies for diversifying and intensifying language learning in order to promote plurilingualism in a pan-European context" (Council of Europe 2001, 4). There is no mention of moving towards European integration of educational policies, rather words celebrate diversity and plurilinguialism.

The Council of Europe 2007 Forum report states in Francis Goullier's introductory summation that the CEFR should be regarded "as a descriptive rather than a standard-setting document" and that "it allows all users to analyse their own situation and to make the choices which they deem most appropriate to their circumstances, while adhering to certain key values." (2007a, 7).

The stated aim is to promote language learning, and yet, the CEFR has come to be viewed and adopted as a standard-setting document with ensuing proliferation of linking scales to the CEFR and semi-official status in some member states. This has happened despite the claim to the opposite, and it is underscored by the development a manual which delineates a linkage procedure to be employed to link tests to the CEFR (Council of Europe 2011). As pointed out by Byram and Parmenter (2012, 5), the grid of levels has become the identifying mark of the CEFR. In Executive summary of the 2007 Survey on the use of the CEFR, it becomes clear that the scales are a major concern in member countries:

Issues related to the use of the common reference levels, such as a need for defining additional sub-levels and the repetitiveness and lack of detail of some descriptors were

indicated as the most acute problems by a majority of countries. (Council of Europe 2007b)

The worries about CEFR levels

Nearly all major testing institutions have developed linkage from their own test scores and levels to those of the CEFR, including IELTS and TOEFL. ALTE (Association of Language Testers in Europe) has developed a framework relating the CEFR levels to testing on a national level for immigrants and others, i.e. high-stakes testing.

These measures indicate a general trust in the CEFR and its level descriptors. The situation is that the CEFR is being referred to as an indicator of levels of qualifications which are required in situations which will influence people's lives. A Danish university for instance lists C2 as an entrance requirement level for language programmes. The latter example is probably in line with the intensions of the Council of Europe to make transfer easier among educational institutions within the member countries.

The worry is that the levels are interpreted and applied in a variety of ways which are not always transparent. One dimension of the problems is attached to the level descriptors themselves. There are two ways at least where interpretation may lead to disparate understandings; one is the interpretation of the formulations of the descriptors themselves, the other is the interpretation of how they can or should be transformed into a testing format. Finally, outside the scope of this paper, there are the issues of testing methodology and underlying theoretical foundations (for extensive writings on this see for example Alderson 2005; and Bachman 1990; Bachman and Palmer 2010).

Assessment study

The increasing pervasiveness of the CEFR and ensuing uneasiness about its usefulness for higher levels of language education in combination with an observation of testing in relation to these levels which seems somewhat problematic constitute the background for the present study and the desire to build a foundation for discussing the CEFR.
The respondents were all trained linguists and highly proficient speakers of English and teachers of English/American literature, history, composition, academic writing, and communication at university level.

The pilot investigated two approaches of assessment and interpretation. The first approach was an assessment of the descriptors for reading and writing. The second was an assessment of an actual piece of writing and its relation to the CEFR descriptors for writing.

The first investigation, in which respondents were asked to interpret the CEFR descriptors for reading and writing, the interest was writing but reading was included for reasons of comparison for consistency of interpretation.

The respondents were asked to place the A1, A2, etc. descriptors in a hierarchy and the results in Table 1 indicate the ability of the respondents to rank descriptors to comply with the ranking of the CEFR grid.

LEVEL	A1	A2	B1	B2	C1	C2	Average
READING	80	80	60	80	20	20	63
WRITING	100	80	80	60	20	40	63
AVERAGE	90	80	70	70	20	30	

Table 1. Ability to interpret CEFR descriptors according to level in per cent.

The level of success in their assessment of the CEFR level descriptors is on average the same for reading as well as writing, with small internal variation; for both a correctness level of 63.33%. The table also illustrates that it is especially the advanced levels of C1 and C2 that respondents find it difficult to assess correctly, and this is true for both reading and writing.

Approximately four weeks after the experiment with the hierarchy assessment, the respondents were asked to assess an actual piece of writing and to align it with one of the descriptor levels for writing in the CEFR. Only one respondent adhered to the straightforward and sharp lines of the CEFR. The other respondents, despite instructions to follow the CEFR grid gave their reply with comments that placed to assessment between the categories. Subsequently, this was transformed and coded into a numerical system somewhat along the lines of the ALTE system. It meant that the grid was coded with numerical points and in-between assessments were given half-points. The results are given in Figure 1.

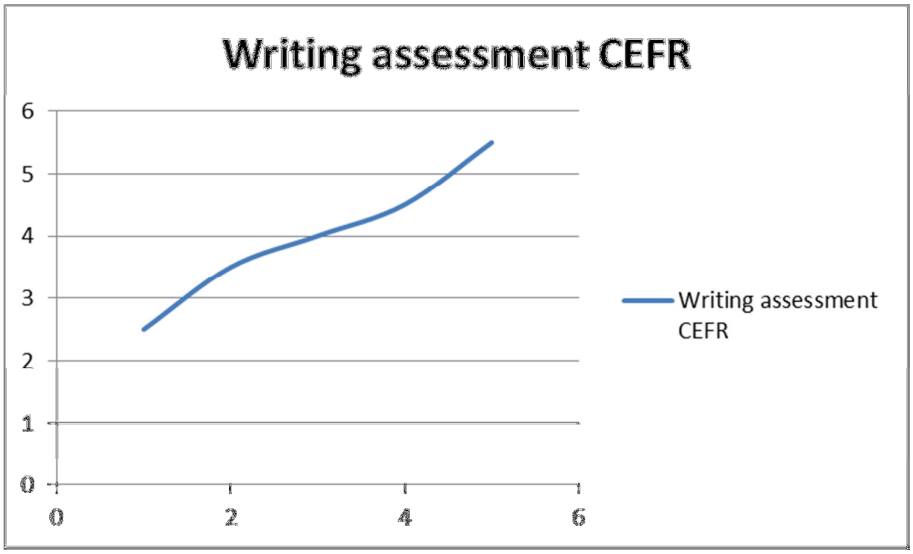


Figure 1. Variation in writing assessment (point scale 1 to 6).

The range in assessments crosses two barriers of break-off points; namely those between Basic User and Independent User on the one hand, and between Independent User and Proficient User on the other hand. This divergent assessment of a given piece of writing is especially worrying if the break-off points are used in high-stakes evaluation, as in the case of university entrance requirements or granting of permits of residence.
The above results find support in other studies. Kiszely and Szabó presented results similar to this study at the EALTA conference in 2009. Their results were documented with thorough statistical analysis and shoved significance levels of $p \leq 0.05$.
Alderson et al. (2004, 2) quoted in Fulcher (2004, 259) warns that "it is far from clear that the abstract statements in the CEFR can be turned into items that illustrate or exemplify the different CEF levels".

Broeder and Martyniuk (2008), without touching on the content knowledge underlying each level, writes about the importance of knowing "what levels language skills are achieved when people learn languages in a formal as well as informal contexts" (209) and goes on to give an overview, which at the national level is illustrated by the French Ministry of Education guidelines. The highest level, the Baccalaureate level, is set as corresponding to "level B2 of the CEFR in the first language studied and B1 in the second language studied" (221).
this example shows how careful one has to be with transnational frameworks which more or less openly stipulate unification or standardisation. The Danish equivalence of the

Baccalaureate would be C2 and C1 respectively; and as mentioned above C2 equivalence is required to enter language programmes at least at one Danish university.

Conclusion

The study results outlined above give grounds for worry on the application of the CEFR, which appears to be expanding at a fast rate. The problem is not the framework per se, but the idea that it can be applied uniformly in disparate contexts and for disparate purposes across 47 nations with differences in educational traditions and educational philosophies, including different levels of teaching and assessment in foreign language learning. It is stated in the Executive summary (Council of Europe 2007b) of the survey on the use of the CEFR at national level in the Council of Europe member states that "The responses show that the CEFR has influenced the development and planning of a number of curricula for primary, secondary, upper secondary schools or adult and higher education." (6).

In the Council of Europe 2007 Forum Report, Daniel Coste (2007) says that: ""The question is not "How do I know that your B2 is my B2?", but "How can I compare my B2 with your B2?" Or, more directly, "What's your B2 like?""(43). However, the study reported on in this paper indicates a need for asking precisely those questions as did Kiszely and Szabó in 2009. Weir (2005) similarly pointed to limitations of the CEFR especially if used prescriptively. In line with Fulcher (2004) this paper concludes that national levels of exam setting and an adherence to evaluation in harmony with educational philosophy prevalent in a given country is the best safeguard for student rights. Evaluations and exams are fraught with problems at the best of times without adding to them by adopting frameworks that are not suited for the purpose.

References

Alderson, C. J. (2005). Diagnosing Foreign Language Proficiency: The interface between learning and assessment. London and New York: Continuum.

Bachman, L. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L. & Palmer, A. (2010). Language Assessment in Practice. Oxford: Oxford University Press.

Broder, P. & Martyniuk, W. (2008). Language Education in Europe: The Common European Framework of Reference. In Encyclopedia of Language and Education, ed. N. van Deusen-Scholl, and N. H. Hornberger, 2nd edition, Vol.4, 209-226. Springer Science+Business Media LLC.

Byram, M. & Parmenter, L. (Ed.). (2012). The Common European Framework of Reference: The Globalisation of Language Policy. Bristol: Multilingual Matters.

Coste, D. (2007). Contextualising uses of the Common European Framework of Reference for Languages. In Council of Europe. Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities. 38-47. Forum, 6-8 February. Strasbourg: Council of Europe. Language Policy Division.
http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage

Council of Europe. (2001). Common European Framework of References for Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.
http://www.coe.int/t/DG4/Portfolio/documents/Framework_EN.pdf
Council of Europe. (2007a). The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and

responsibilities. Forum, 6-8 February. Strasbourg: Council of Europe. Language Policy Devision.
http://www.coe.int/t/dg4/linguistic/Forum07_webdocs_EN.asp#TopOfPage

Council of Europe. (2007b). Executive summary of results of a survey on The use of the CEFR at national level in the Council of Europe Member States. Strasbourg. Council of Europe Language Policy Division.

Fulcher, G. (2004). Deluded by Artifices? The Common European Framework and Harmonization. Language Assessment Quaterly, 1, (4), 253-266.

Kiszely, Z. & Szabó, G. (2010). Is My B2 the Same as Your B1? Comparing Language Examinations' Levels. Presented at EALTA Conference. Retrieved from http://www.ealta.eu.org/conference/2009/docs/posters/Poster_Kizely_Szabo.pdf

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. Language Testing, 22, (3), 281-300.

Elena Volodina & Sofie Johansson Kokkinakis

University of Gothenburg, Gothenburg, Sweden

elena.volodina@svenska.gu.se - sofie.johansson.kokkinakis@svenska.gu.se

Compiling a Corpus of CEFR-Related Texts

Bio data

Elena Volodina received PhD in Linguistics in Moscow, Russia (1998) and MA in Computational Linguistics in Gothenburg, Sweden (2008). She works as a research engineer at Gothenburg University since 2010, her primary interests being in Intelligent Computer-Assisted Language Learning, text and sentence readability, Corpus Linguistics.

Sofie Johansson Kokkinakis received PhD in Language Technology (2005) and MA in Computational Linguistics, both at the University of Gothenburg, Sweden. She currently works as a Researcher in language technology and the Director of the Institute for Swedish as a Second Language, Department of Swedish, University of Gothenburg. Her interests comprise computer-based lexical and content related analysis of texts (in particular secondary school text books and student written texts), lexical profiling and readability, genre specific texts and academic language use, morpho-syntactic and semantic analysis, computer-based assessment of language skills, corpus linguistics and learner corpora, and intelligent computer-assisted language learning

Abstract

This paper reports on initial efforts to compile a corpus of course book texts used for teaching CEFR-based courses of Swedish to adult immigrants. The research agenda behind compiling such a corpus comprises the study of normative "input" texts that can reveal a number of facts about what is being taught in terms of explicit grammar, receptive vocabulary, text and sentence readability; as well as build insights into linguistic characteristics of normative texts which can help anticipate learner performance in terms of active vocabulary, grammatical competence, etc. in classroom and testing settings.

The CEFR "can-do" statements are known to offer flexibility in interpreting them for different languages and target groups. However, they are nonspecific and therefore it is difficult to associate different kinds of competences and levels of accuracy learners need in order to perform the communicative tasks with the different CEFR levels. To address this problem a systematic study needs to be performed for each individual language, both for "input" normative texts and "output" learner-produced texts. In this project we take the first step to collect and study normative texts for Swedish.

The article describes the process of corpus compilation, annotation scheme of CEFR-relevant parameters, and methods proposed for text analysis, namely statistic and empiric methods, as well as techniques coming from computational linguistics/machine learning.

Short paper

Introduction

Since the acceptance of Common European Framework of References for Languages (CEFR) in 2001 (Council of Europe, 2001) many countries inside and outside Europe have

abandoned previous practices in language teaching and assessment in favour of the CEFR. The CEFR scale, consisting of 6 proficiency levels, is described intentionally vaguely to cater for the diversity of different languages. As a consequence, there are voices among researchers and educators demanding explicit interpretation of each proficiency level for each individual language in terms of required vocabulary scope, grammatical competence, etc. (Byrnes 2007; Little 2007; Little 2011; Milton 2009; North 2007; Westhoff 2007).

It is known to be rather controversial to break down the CEFR "can-do" statements into concrete constituents, partly due to the "human factor". Course material producers and teachers often go by their subjective "expert judgements" and intuitions, not necessarily agreeing with each other. However, we take it for granted that teachers' interpretations of CEFR guidelines, subjective when taken individually, present an objective ground for generalizations and approximations about language complexity and level-wise content, when taken collectively. Thus, we assume that, given texts used for CEFR-based courses, we can perform empiric studies of a number of linguistic aspects expected of learners at different levels, for example vocabulary scope, most common grammar per level, text complexity, sentence complexity. Apart from that, we are interested in studying typical linguistic features for texts of different CEFR-based themes (topical domains).

Background

Texts related to CEFR-based language learning fall into two categories as shown in figure 1: (1) "input" or normative texts provided by course book writers or selected by teachers; and (2) "output" or learner produced texts showing learner performance at the studied level.

Figure 1. Texts in L2 context

The study of learner produced language is a large and active area of research in second language learning (Johansson Kokkinakis & Magnusson, 2011; Hultman & Westman, 1977; Nyström, 2000; Östlund-Stjärnegårdh, 2002). In Sweden, as far as we know, most research in this area is conducted with respect to language development theories, such as "the processability theory" (Pienemann 1998). However, since CEFR is widely spread in everyday practice, there is a need for CEFR-based analysis of learner language as well. Examples of projects devoted to CEFR-based studies of learner-produced language for other languages than Swedish are given under the SLATE research network (Carlsten, 2012; Hawkins & Buttery, 2009;more under <http://www.slate.eu.org/>).

In contrast to research within learner-produced language, we are not aware of any active studies performed on normative texts used in CEFR based courses or on correlation between normative texts and learner production, in spite of the fact that teachers, researchers and language assessors keep expressing the need for formalizing CEFR descriptors in terms of concrete grammar and vocabulary syllabus. In the project described in this paper we aim at collecting normative texts to fill in the gap and to form the ground for that kind of studies for Swedish.

Why compile this corpus?

Given the availability of electronic resources of the above-mentioned kind, we can engage in a number of important and interesting from the research point of view studies, eventually useful even outside research circle. For example, using a combination of statistic and empirical methods, as well as methods derived from computational linguistics (e.g. machine learning), we can study features characteristic of different CEFR levels. The possible outputs of such studies comprise (1) an instrument for automatic classification of texts by CEFR level based on text readability; (2) an instrument for automatic classification of sentences by CEFR level based on sentence readability; (3) an instrument for automatic classification of texts into thematic domains based on machine learning approach; (4) receptive vocabulary scope per proficiency level; and (5) receptive grammar scope per level. The main question are, then:

- which linguistic aspects are most important at each particular CEFR level, and why (at sentence and text levels individually); how the identified linguistic aspects match the "can-do" descriptors;
- which linguistic features are characteristic of texts of different thematic domains; and how such texts can be automatically identified;
- which words and how many per proficiency level are important to learn;
- which grammar students are most exposed to during the course of studies.

The studies based on the corpus may help us answer some of the questions often raised in the CEFR-based language testing context, for example, if there is a link between 'can do' performance statements and areas of linguistic knowledge; or to what extent the levels can be made more explicit in terms of required vocabulary and grammar. We view our study as an evidence-based interpretation of vague CEFR descriptors for different levels into concrete linguistic constituents based on expert interpretations of many experienced teachers and course book writiers.

Corpus compilation: first experiences

Interviews with teachers and publishers

Course materials are often copyrighted by publishing houses and cannot be freely obtained, to say nothing of being freely distributed in electronic format. To identify relevant course materials, a number of teachers of CEFR-related courses have been interviewed. It has turned out the materials available in the form of course books only in few cases implicitly mention European framework (Levy Scherrer & Lindemalm, 2008; Levy Scherrer & Lindemalm, 2009; Göransson & Parada, 2010; Göransson et al., 2010; Folkuniversitetets förlang, 2007; Åström, 2011; Åström, 2012; Trevisani, 2011); whereas a number of course books do not provide any indication what level according to the European scale of references they are aimed at, but have been interpretered by teachers as appropriate ones at certain levels (Holm et al., 2001; Ballardini, 2001a, 2001b; Risérus et al., 2002;).

All the relevant publishers have subsequently been contacted for electronic materials. However, texts in electronic format have proven to be rather difficult to obtain. Of all the contacted publishers only Liber has shown understanding and provided files for our research. To tackle the problem of lacking texts, we opted for an optical scanning approach. The total amount of coursebooks in pages is 3187; which corresponds to an estimated size of approximately 3 mln tokens.

Optical scanning and its challenges

We have subcontracted scanning to a relevant digital centre. Our "pilot" level has become B1, with 3 different coursebooks, each containing mixed contents (e.g. half the book B1 and half the book B2; or a part of the book A1/A2, the rest B1), totalling 565 pages.
During this stage we have encountered a number of challenges. Without getting in to the details of digital document analysis or techniques for optical text recognition, (which is a separate research ares, see for example International Journal on Digital Analysis and

Recognition <http://www.springer.com/computer/image+processing/journal/10032>) we describe here what we have encountered practically.

First of all, book availability has caused some problems. Since some of the books are rather old, e.g. from 2001, they are neither sold via book stores, nor are they available at the library. In some cases we could find copies from teaching staff, but often they contained scribbles that interfered with scanning.

Second, it is a challenge to scan correctly texts that are diagonally placed, as in figure 2.

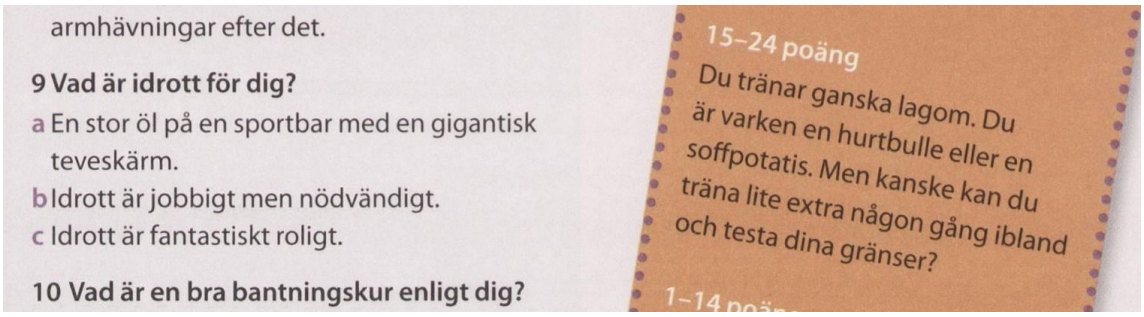


Figure 2. Example of a diagonally placed text.

The extracted text from the scanned document looked like that:

```
9          Vad är idrott för dig?
a En stor öl på en sportbar med en gigantisk teveskärm.
b Idrott är jobbigt men nödvändigt.
c Idrott är fantastiskt roligt.
braät"öTn!hU"bu"e!Det
ar
^ attrÖraPåsigoch mer> kanske* CL. . a
tänka på att
danske ska r,
: Osanna . Uanfa
; Kr°PPenbehöv ""ani"9iliVet ; 'bland.      0också vila
:
ar varken enh5ka 'd90m Du SOffP«afe.Men?U"ee"eren 'rän*l*eeZ
.kansk^ndu
• 0ch^«nagZ2fn9ibl*™
10 Vad är en bra bantningskur enligt dig?
```

Figure 3. Result of an optical scanning of a diagonally placed text.

Starting with line 5 and till the last line but one (figure 3) there is a lot gibberish. Some of the words or phrases coincide with the phrases in the diagonally placed text but very inconsistently.

Texts given in several blocks or tables (as shown in figure 4) present a problem of texts interfering with each other (figure 5).

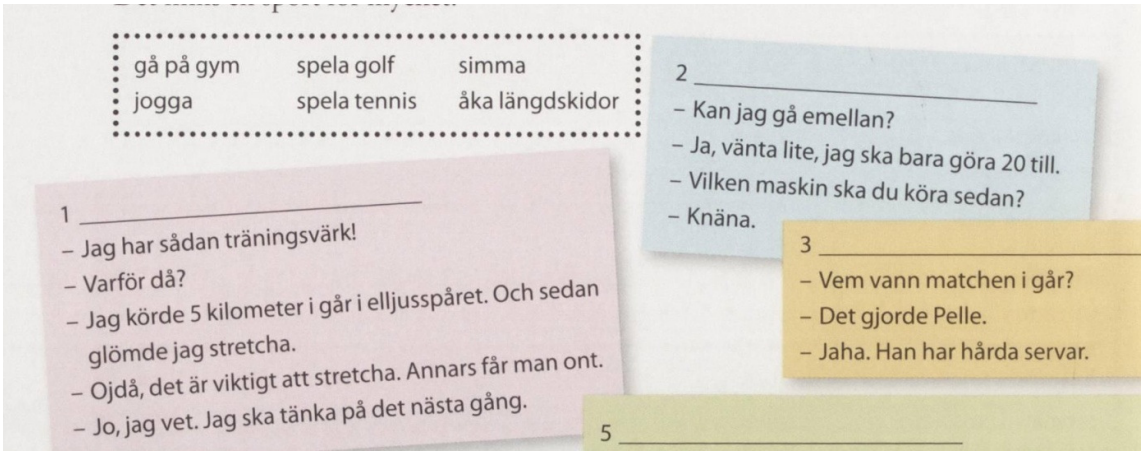


Figure 4. A view of several texts placed in the form of "table" and therefore setting risk to have texts bumping into each other.

1 ga pa gym jogga
2 spela golf spela tennis
3 simma
4 åka längdskidor
5 -jag har sådan träningsvärk!
6 Kan jag gå emellan?
7 -la, vänta lite, jag ska bara göra 20 till.
8 - Vilken maskin ska du köra sedan?
9 - Knäna.
10 3_
11 _ Varför då?
12 - Jag körde 5 kilometer
13 glömde jag stretcha.
14 - Ojdå, det är viktigt att
15 går i elljusspåret. Och sedan
16 stretcha. Annars får man ont.
17 - Vem vann matchen i går?
18 - Det gjorde Pelle.
19 - Jaha. Han har hårda servar.
20 - Jo, jag
21 vet. Jag ska tänka på det nästa gang.

Figure 5. Result of optical scanning of text presented in figure 4.
Line numbers added for easier interpretation.

As can be seen from figure 5, lines 1-4 represent the word list in figure 4; line 5 starts dialogue nr.1, whereas lines 6-9 refer to dialogue nr.2, lines 11-16 continue dialogue nr.1, though in a scrambled order. The correct order should be (given here in line numbers): 11, 12, 15, 13, 14, 16.

We made a decision to ignore texts that haven't been correctly scanned unless it demands little effort to restore the correct text. We have therefore lost a bit of text mass during the post-scanning step.

Annotation
Coursebook texts annotation consists of two steps:

1. annotation for CEFR-relevant variables and
2. annotation for linguistic parameters.

Annotation for CEFR variables

We used Lärka, the ICALL[1] platform for Swedish (Volodina & Borin 2012), as the basis for the editor. Figure 6 presentes the course book editor view:

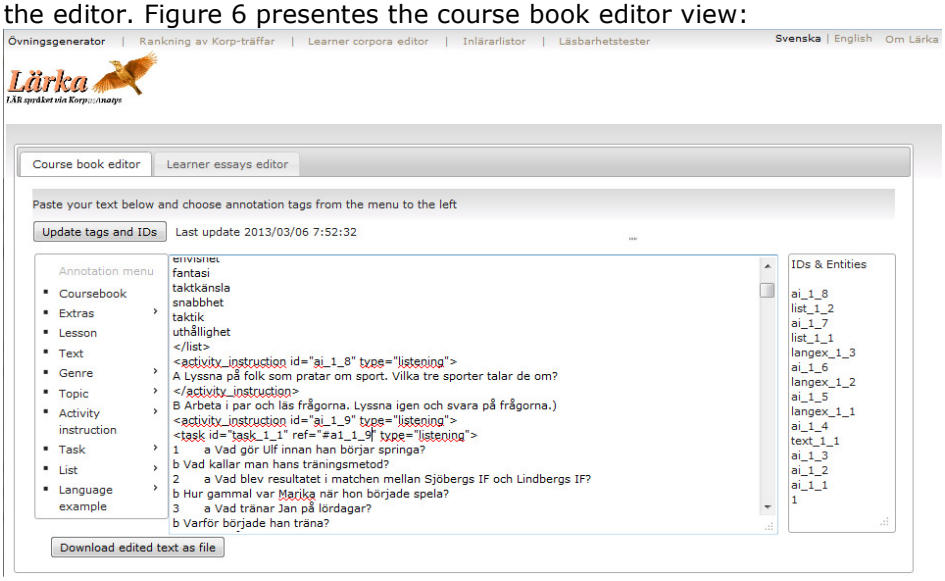


Figure 6. Course book editor view in Lärka

The menu on the left inserts different tags into the text field; the field on the right keeps track of the ids used throughout the file.

The most interesting from CEFR point of view is the taxonomy of text variables. We have divided the text mass in course books into "extras" (foreword, contents, acknowledgements, etc.) and lessons (i.e. chapters). Lessons contain different types of learner-interesting language that we have divided into texts, activity instructions, tasks, lists and language examples. A more fine-grained division is shown in figure 7.

Text genres is a modified version of genre families described in Martin & Rose (2008).
It has been extended by some macrofunctions as described in the CEFR guidelines, e.g. exposition, exegesis (Council of Europe, 2001, p.126); as well by the genre marked as "other" which contains text types that we couldn't place in any of the main three categories (narration, facts, evaluation). Among the a-typical text types are puzzles, rhymes, lyrics, questionnaires, letters, etc. The genre taxonomy is not final since we expect to meet other deviating categories during the annotation work

1 ICALL – Intelligent Computer-Assisted Language Learning

Text parameters: Genre	Text parameters: Topic	Other types of text in lessons

Genre
• Narration
  • Personal story
  • Fiction
  • Description
  • News article
• Facts
  • Historical facts
  • Biography
  • Autobiography
  • Explanation
  • Instruction
  • Rules
  • Procedures
  • Report
  • Demonstration
• Evaluation
  • Argumentation
  • Exposition
  • Discussion
  • Personal reflection
  • Review
  • Interpretation, exegesis
  • Persuasion
• Other
  • Dialogue
  • Puzzle
  • Rhyme
  • Lyrics
  • Questionnaire
  • Letter
  • Language tip
	Topic
	• Personal identification
	• House and home, environment
	• Daily life
	• Free time, entertainment
	• Travel
	• Relations with other people
	• Health and body care
	• Education
	• Shopping
	• Food and drink
	• Services
	• Places
	• Languages
	• Weather
		Activity instruction
		• Listening
		• Reading
		• Writing
		• Speaking
		• Discussion
		• Grammar exercise
		• Vocabulary exercise
		• Text question
		Task
		• Listening
		• Reading
		• Writing
		• Speaking
		• Discussion
		• Grammar exercise
		• Vocabulary exercise
		• Text question
		• Gaps
		List
		• Vocabulary
		• Grammar
		• Sentences
		Language example
		• Vocabulary
		• Grammar
		• Pronunciation
		• Spelling
		• Writing

Figure 7. Submenus of the main annotation menu for genre, topic, activity instruction, task, list and language example

Topics have also been taken from the CEFR document (Council of Europe, 2001, p.52). As in the case with the genres, we expect the list of topics to grow during the annotation period to cover the diversity of text topics in the course books.

The division of the language used in lessons into texts and other categories is made to cater for different types of research that can be performed once the corpus is available.

Once the course book editor is stable, it will be available for use for any other L2 language course books annotation, language independent. Since it is web-based, it can be accessed from anywhere without prior installation.

Annotation for linguistic variables

Annotation for linguistic variables includes annotation for parts of speech (pos in figure 8), morpho-syntactic information (msd), syntactic relations (ref, dephead, deprel), lemmas, and linking to morphology lexicon (lex, saldo). This is an automated procedure that is used in Korp[2] import pipeline (Borin et al. 2012). Example of how a text can look after this annotation is given in figure 8.

2 Korp – an infrastructure for storing and browsing a large collection of Swedish texts (Borin et al. 2012); www.spraakbanken.gu.se/korp

```
<w pos="PS" msd="PS.UTR.SIN.DEF" lemma="|jag|" lex="|jag..pn.1|" saldo="|jag..1|"
prefix="|" suffix="|" ref="01" dephead="02" deprel="DT">Min</w>

<w pos="NN" msd="NN.UTR.SIN.IND.NOM" lemma="|kompis|" lex="|kompis..nn.1|"
saldo="|kompis..1|" prefix="|komp..nn.1|kompa..vb.1|" suffix="|is..nn.1|" ref="02"
dephead="03" deprel="MS">kompis</w>

<w pos="KN" msd="KN" lemma="|och|" lex="|och..kn.1|" saldo="|och..1|" prefix="|"
suffix="|" ref="03" dephead="" deprel="ROOT">och</w>

<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|jag|" lex="|jag..pn.1|" saldo="|
jag..1|" prefix="|" suffix="|" ref="04" dephead="05" deprel="SS">jag</w>

<w pos="VB" msd="VB.PRS.AKT" lemma="|ha|" lex="|ha..vb.1|" saldo="|ha..1|"
prefix="|" suffix="|" ref="05" dephead="03" deprel="MS">har</w>

<w pos="VB" msd="VB.SUP.AKT" lemma="|känna|" lex="|känna..vb.2|känna..vb.1|"
saldo="|känna..2|känna..4|känna..1|känna..3|" prefix="|" suffix="|" ref="06"
dephead="05" deprel="VG">känt</w>
```

Figure 8. Example of annotation for linguistic variables

Intended corpus use

Special efforts have been undertaken to interpret CEFR guidelines as sets of Reference Level Descriptions[3] as well as to establish procedures to relate language exams to the CEFR (Council of Europe, 2009; Khalifa et al., 2010; Szabó, 2010; Dávid, 2010; Jones et al., 2010), but to the best of our knowledge that has not yet been done for Swedish.

The availability of electronic resources of the described type opens an opportunity to engage in an evidence-based interpretation of the CEFR descriptors. "Evidence-based" in the context of this project is understood as course book materials collected into a linguistically annotated corpus. They present an evidence of conscience expert interpretations of CEFR guidelines into concrete samples of teaching material.

To address the problem of non-specificity of the CEFR descriptors for different levels of language proficiency, a systematic study needs to be performed for each individual language, both for "input" normative texts and "output" learner-produced texts. Attempts at aligning texts and tests with CEFR are ongoing (Khalifa et al., 2010; Szabo, 2010; David, 2010; Jones et al., 2010) with what could be called a top-down approach, i.e. starting from CEFR descriptors and going all the way down to the actual selection of appropriate texts/language samples, interpreting the CEFR descriptors on the way. This process consists of four procedures according to the Manual (Council of Europe, 2009): familiarization, specification, standardization, and empiric validation. We suggest a bottom-up approach, where we start from the actual language samples labeled for levels, i.e. preselected reading materials for different levels, analyze them for linguistic constituents with the help of machine learning algorithms and then try to map the identified constituents to the CEFR descriptors. The two approaches should be viewed as complementary rather than exclusive of each other.

Once ready, the collection of normative texts introducd in section 3 can be studied internally to generate an instrument that can reliably classify any arbitrary Swedish text by its appropriate CEFR level and domain. Availability of the corpus will also make it possible to identify receptive vocabulary and grammar scope per proficiency level.

3 http://www.coe.int/t/dg4/linguistic/dnr_en.asp

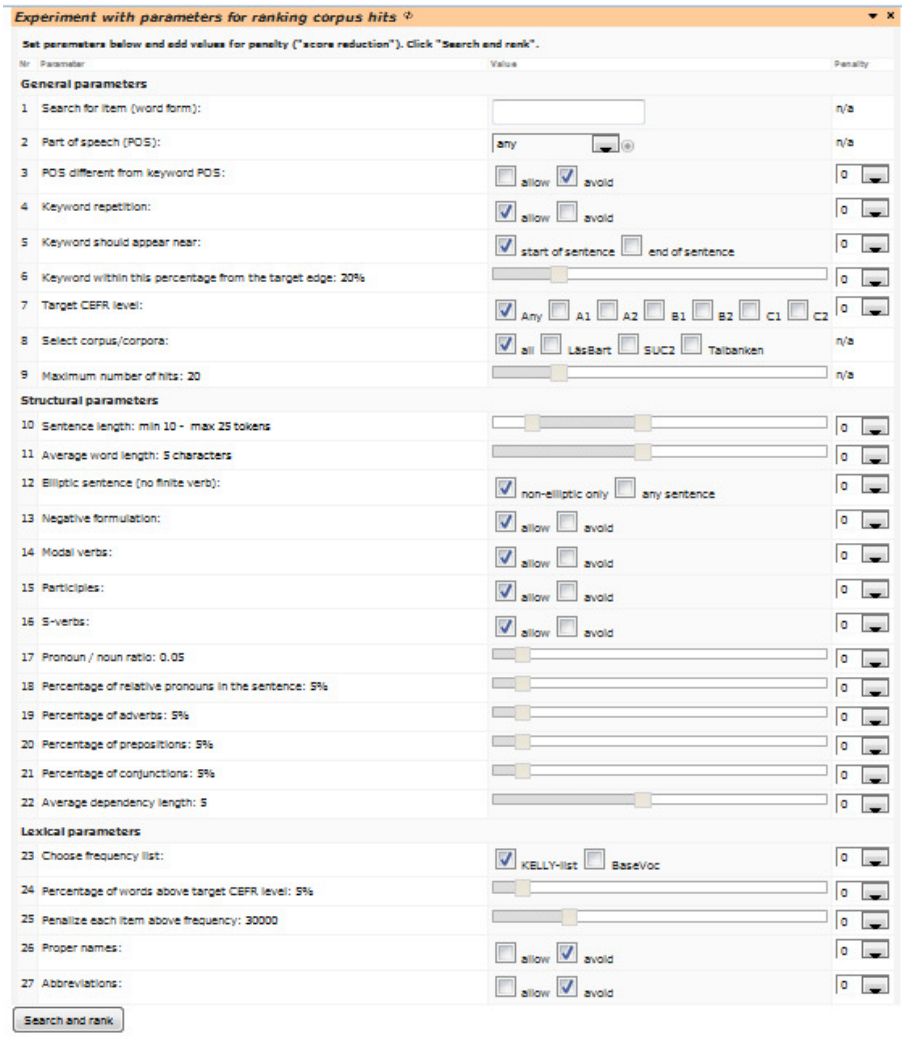


Figure 9. Linguistic parameters for sentence and text analysis

The first use of the corpus is planned for an internal project that will help us identify an automatic approach to the readability assessment of Swedish sentences in the L2 context (Pilan et al., forthcoming). More concretely, the aim is to create an algorithm which would try to predict at which language learning level students are able to understand sentences containing certain lexical, syntactical, morphological and other linguistic elements. This approach is a combination of evidence-based empiric methods combined with statistical and machine-learning techniques and leads us to the explicit mapping between required vocabulary, grammar and syntax and the reached CEFR levels; the identified linguistic parameters can be further connected to the level-wise 'can-do' statements.

The linguistic parameters we have selected so far for scrutiny are presented in the left column of figure 9. We initially plan to study A1, A2, B1 and B2 course book texts in contrast to non-restricted texts used for native speakers coming from generic balanced corpora of Swedish. This will show us how the linguistic features in figure 9 are distributed in normative texts of different proficiency levels.

The same type of study is planned for text-long contexts at different levels.

As a further step we intend to collect a corpus of student essays written at different CEFR levels and compare linguistic features used in normative texts, i.e. the ones that learners are expected to cope with receptively when using course books, versus learner-produced texts, showing how these features are reflected in their productive use.

Concluding remarks

In this paper we have presented our initial work on compiling and annotating a corpus of CEFR-based course book texts, and outlined the prospects of its usage for CEFR-based pedagogical studies. This kind of data labeled for CEFr levels, topical themes, etc. is critical for pedagogical empirical studies like the ones proposed above since it facilitates conclusions, generalizations and approximations about language use in L2 context. With this project, we lay the ground for further pedagogically relevant studies of CEFR related texts in Swedish.

Acknowledgements

The project presented here has been financed partly by the Swedish Department at the university of Gothenburg (UGOT) and partly by Språkbanken, UGOT.

We extend our thanks to the publishing house Liber for providing electronic materials for our research.

References

Ballardini, K., Stjärnlöf, S. & Viberg, Å. (2001a). Nya Mål: svenska som andraspråk. 1. Natur & Kultur.

Ballardini, K., Stjärnlöf, S. & Viberg, Å. (2001b). Nya Mål 2 Lärobok. Natur & Kultur.

Borin, L., Forsberg, M. & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. Proceedings of LREC 2012. Istanbul: ELRA, p.474–478.

Byrnes H. (2007). Perspectives. The Modern Language Journal, 91, 641–645.

Carlsten, C. (2012). Proficiency Level – a Fuzzy Variable in Computer Learner Corpora. Applied Linguistics, 33(2), 161-183.

Council of Europe. (2001). The Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Council of Europe. (2009). Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR). A Manual, Strasbourg: Language Policy Division.

Dávid, G. A. (2010). Linking the general English suite of Euro Examinations to the CEFR: a case study report. In Martyniuk, W. (Ed.) Aligning Tests with the CEFR. Cambridge University Press, p.177-203.

Folkuniversitetets förlag. (2007). På svenska! 3 lärobook. Folkuniversitetets förlag.

Göransson, U. & Parada, M. (2010). På svenska! 1 lärobok. Folkuniversitetets förlag.

Göransson, U., Helander, A. & Parada, M. (2010). På svenska! 2 lärobok. Folkuniversitetets förlag.

Hawkins, J. A. & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In Taylor, L. & Weir, C.

J.(Eds). Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment, 158-175. Cambridge: Cambridge University Press.

Holm, B., Nyborg, R. & Pettersson N.O. (2001). Svenska utifrån: Lärobok i svenska. Svenska institutet.

Hultman, T. G. & Westman, M. (1977). Gymnasistsvenska. Lund: Liber Läromedel.

Khalifa, H., French, A. & Salamoura, A. (2010). Maintaining alighnment to the CEFR: the FCE case study. In Martyniuk, W. (Ed.) Aligning Tests with the CEFR. Cambridge University Press, p.80-101.

Johansson Kokkinakis, S. & Magnusson, U. (2011). Computer based quantitative methods applied to first and second language student writing. Young urban Swedish. Variation and change in multilingual settings.University of Gothenburg. 105-124.

Jones, N., Ashton, K. & Walker, T. (2010). Asset Languages: a case study of piloting the CEFR Manual. In Martyniuk, W. (Ed.) Aligning Tests with the CEFR. Cambridge University Press, p.227-246.

Levy Scherrer, P. & Lindemalm, K. (2008). Rivstart: A1+A2 Textbok. Natur & Kultur.

Levy Scherrer, P. & Lindemalm, K. (2009). Rivstart: B1+B2 Textbok. Natur & Kultur.

Little D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. The Modern Language Journal 91, 645–655.

Little D. (2011). The Common European Framework of Reference for Languages: A research agenda. Language Teaching, 44 (3), 381-393.

Martin, J.R. & Rose, D. (2008). Genre Relations. Equinox Publishing Ltd.

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. Toronto: Multilingual Matters.

North, B. (2007). The CEFR illustrative descriptor scales. The Modern Language Journal, 91, 656–659.

Nyström, C. (2000). Gymnasisters skrivande. En studie av genre, textstruktur och sammanhang. Uppsala: Uppsala universitet.

Pienemann, M. (1998). Language Processing and Second Language Development: Processability Theory. Amsterdam: John Benjamins.

Pilan, I., Volodina, E. & Johansson, R. (forthcoming). Automatic selection of optimal sentences for language learning exercises. Proceedings of EuroCALL 2013.

Risérus, H., Sandahl, Å. & Stjärnlöf, S. (2002). Nya mål 3. Natur & Kultur.

Szabó, G. (2010). Relating language examinations to the CEFR: ECL as a case study. In Martyniuk, W. (Ed.) Aligning Tests with the CEFR. Cambridge University Press, p.133-144.

Trevisani, M. (2011). Avancera skriv: skrivövningar. Liber.

Volodina, E. & Borin, L. (2012). Developing a freely available web-based exercise generator for Swedish. CALL: Using, Learning, Knowing. EuroCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings. Eds. Linda Bradley and Sylvie Thouësny. Research-publishing.net, Dublin, Ireland

Westhoff G. (2007). Challengens and Opportunities of the CEFR for Reimagining Foreign Language Pedagogy. The Modern Language Journal 91, p.676–679.

Åström, M. (2011). Språkporten Bas. Studentlitteratur AB.

Åström, M. (2012). Språkporten 1, 2, 3 Textbok med webbdel: Svenska som andraspråk. Studentlitteratur AB.

Östlund-Stjärnegårdh, E. (2002). Godkänd i svenska? Bedömning och analys av gymnasieelevers texter. Uppsala: Uppsala universitet. Carlsten, C. (2012). Proficiency Level – a Fuzzy Variable in Computer Learner Corpora. Applied Linguistics, Volume 33(2), p.161-183.

Ying Zheng & John De Jong

Pearson, London, United Kingdom

ying.zheng@pearson.com - john.dejong@pearson.com

Linking to the CEFR: Validation Using Priori and Posteriori Evidence

Bio data

Ying Zheng (Ph.D.), Head of Psychometrics and Research, ELT, Pearson. Ying joined Pearson in 2009. Her research interests include psychometric analysis of language testing data, English as second/foreign language learner characteristics, and quantitative research methodology. Ying's publications can be found in Language Testing, Canadian Journal of Education, Canadian Modern Language Review, International Journal of Pedagogies & Learning, and TESL-EJ. Her area of work covers language tests from China, Canada, and the UK.

Abstract

Linking tests to international standards such as the CEFR is a way of establishing criterion-referenced validity. As is widely acknowledged, validation is a continuous process of quality monitoring. In addition to a posteriori validity evidence, a priori validity evidence - such as test design decisions and the evidence that supports these decisions - also makes a significant contribution to the establishment of validity (Schilling, 2004).

This paper reports on how CEFR scales are operationalized in practice in the course of developing an international English test. Measures to link the test to the CEFR have been studied at different stages of test development. The measures include activities that incorporate the use of CEFR scales in item writing, rating scale developing and human rater training.

A posteriori statistical evidence has been collected from both field tests and live tests. Field test data were used to establish the extent to which scores from this test can be linked to the CEFR, which involved both a test taker-centred approach and an item-centred approach. For the test taker-centred approach, test taker responses on five items from three item types were used: Writing essay (one item), Oral description of an image (two items) and Oral summary of a lecture (two items). These responses were rated on the relevant CEFR scales for writing and speaking by two human raters, independently of the ratings produced to score the test. For the item-centred approach, item writers were required to indicate the most appropriate CEFR level of ability for each item. These estimates were compared with the average item difficulty obtained from field tests.

Furthermore, this paper also reports on the ongoing item seeding process, whereby new test items are seeded in live tests and, following analysis of the results, benchmarked to CEFR-referenced item difficulties.

Short paper

Research Background

Linking tests to international standards such as the Common European Framework of Reference for Languages (CEFR) is a way of establishing criterion-referenced validity. As is widely acknowledged, validation is a continuous process of quality monitoring. In addition to data based statistical analysis, i.e., a posteriori validity evidence, of how a test can claim its alignment to the CEFR, a priori validity evidence - such as test design decisions and the evidence that supports these decisions - also makes a significant contribution to the establishment of validity (Schilling, 2004).

Pearson Test of English Academic (PTE Academic) measures English language proficiency for communication in tertiary level academic settings. It is targeted at intermediate to advanced English language learners. In order to claim that PTE Academic is fit for its purpose, a variety of validity evidence has been collected from the various stages of test development through to its administration. The constructs measured in PTE Academic are the communicative language skills needed for reception, production and interaction in both oral and written modes, as these skills are considered necessary to successfully follow courses and to actively participate in the targeted tertiary level education environment.

This paper reports on how CEFR scales are operationalized in practice in the course of developing PTE Academic. The CEFR describes what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop to be able to act effectively. Language ability is described with CEFR as a number of scales, which include a global scale, skill specific scales, and linguistic competency scales. In the context of PTE Academic, measures to link the test to the CEFR have been studied at different stages of test development. A priori measures include activities that incorporate the use of CEFR scales in item writing. A posteriori evidence includes the statistical validation procedures used to establish the extent to which PTE Academic scores can be linked to the CEFR.

A Priori Validation

Since test scores of PTE Academic are used for university admission purposes, the high-stakes nature of the decisions require this test to be valid for the inferences the test users make, that is, whether test takers have adequate English proficiency to succeed in English-medium tertiary settings. In developing valid test items, quality assurance measures were adopted at each stage of the test development processes.

Qualified item writers are trained to become familiar with two essentail test develolopment documents, i.e., Test Specification and Item Writer Guidelines. Test Specification serves as an operational definition of the constructs the test intends to assess. Item Writer Guidelines includes detailed test specification of PTE Academic and the CEFR scales, which further specified in detail the characteristics of each item and gave item writers rules and checklists to ensure that test items are fit for purpose and suitable for inclusion in the item bank.

In developing reading and listening items, item writers are largely trained in three aspects: 1) target language use situation; 2) selecting appropriate reading or listening texts; and 3) the CEFR scale on reading and listening. The Guidelines explains the characteristics of reading and listening passages through which test takers can best demonstrate their abilities. For the reading items, this includes test sources, authenticity, discourse type, topic, domain, text length and cultural suitability. For the listening items, it includes text sources, authenticity, discourse type, domain, topic, text length, accent, text speed, how often the material will be played, text difficulty, and cultural suitability.

In the section of developing speaking and writing items, the Guidelines explain target language use situation with details of the CEFR scale from levels B1 to C2. In the Guidelines for writing, the purpose of writing discourse and the cognitive process of academic writing are presented in a matrix format with recommendations for preferred item types. The purposes of writing tasks are defined as 1) to reproduce, 2) to organize or reorganize, and 3) to invent or generate ideas. Three types of cognitive processing are differentiated: to learn; to inform; and to convince or persuade. In the Guidelines for speaking, item writers are instructed to produce topics focusing on academic interests and university student life. A list of primary speaking abilities is also provided, including the ability to comprehend information and deliver such information orally, and the ability to interact with ease in different situations.

Writing to the CEFR Levels

This section describes specific procedures involved in the writing of the test items to the CEFR levels. Item writers are instructed to write items with a difficulty level from B1 to C2 on the CEFR scale. Their predictions of item difficulty level was empirically validated when the items were analyzed either though field testing or through live item seeding process. Table 1 presents an overview of the four main stages in the CEFR familiarization trainings for item writers.

Main Stages	**Details**
STAGE 1: Familiar with the definitions of some basic terms used in CEFR	For example: general language competence, communicative language competence, context, conditions and constraints, language activities, language processes, texts, themes, domains, strategies, tasks
STAGE 2: Familiar with the common reference level: the global descriptors	**Proficient user (C2 & C1):** precision and ease with the language, naturalness, use of idiomatic expressions and colloquialisms, language used fluently and almost effortlessly, little obvious searching for expression, smoothly flowing, well-structured language **Independent user (B2 & B1):** effective argument, holding one's own, awareness of errors, correcting oneself, maintains interaction and gets across intended meaning, copes flexibly with problems in everyday life **Basic user (A2 & A1):** interacts socially, simple transactions in shops, etc skills uneven, interacts in a simple way
STAGE 3: Familiar with the sub-scales for four skills	CEFR Overall Written Production and sub-scales CEFR Overall Speaking Production and sub-scales CEFR Overall Listening Production and sub-scales CEFR Overall Reading Production and sub-scales
STAGE 4: Rating candidates' performances based on CEFR	Rate individually Express reasons and discuss with colleagues Compare rating with experts' marks

Table 1: Item Writer CEFR Training Stages

As shown in Table 1, there are four stages in the CEFR familiarization training. The first stage covers the instruction of some key terms that are used in the CEFR descriptors, aiming to facilitate item writer trainees to understand the CEFR in general. By introducing the global descriptors at each level, the second stage give trainees an idea of what kind of tasks and how well the test takers are expected to perform at different levels. The third stage provides the trainees with more detailed descriptions of 'can do' statements. Finally, after becoming familiar with the CEFR scales, item writers are asked to rate several recordings of speaking performances individually, discuss with their colleagues

what ratings they gave and compare their scores and reasons with those given by experts. Table 2 shows an example of CEFR overall written production and subscales.

	CEFR Overall Written Production	**CEFR Writing sub-scales**
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader find significant points.	Creative writing Reports and essays Overall written interaction Correspondence Notes, messages and forms Note taking Processing text Orthographical control Thematic development Coherence and cohesion
C1	Can write clear, well-structured texts on complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples and rounding off with an appropriate conclusion.	
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesizing and evaluating information and arguments from a number of sources.	
B1	Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements in a linear sequence.	
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.	
A1	Can write simple isolated phrases and sentences.	

Table 2: An example of CEFR Overall Written Production and sub-scales

In summary, in the context of PTE Academic, the concept of CEF is built into the essential test development documents and implemented at the initial stage of test item writing. Each item was written with an item writer CEF estimate, which are then cross-validated using statistical evidence.

A Posteriori Validation

This section reports the statistical validation procedures used to establish the extent to which PTE Academic scores can be linked to the CEFR. Statistical procedures for relating PTE Academic scores to the levels of the CEFR scales involved both a test taker-centered approach and an item-centered approach.

A Test Taker-Centered Approach

For the test taker-centered approach, test taker responses on five items from three item types were used: Writing essay (one item); Oral description of an image (two items); and Oral summary of a lecture (two items). Writing essay has 11 scores categories (0-10 points); Oral description of an image has 8 score categories (0-7 points), and Oral summary of a lecture has 5 score categories (0-4 points). These responses were rated on the relevant CEFR scales for writing and speaking by two human raters, independently of the ratings produced to score the test. Given the probabilistic and continuous nature of the CEFR scale, adjacent scores were expected in the model.

The relation between ability estimates based on scored responses on the above PTE Academic test items and the CEFR is displayed in Figure 1, with one for the written responses, and the other for the oral responses. The horizontal axis ranges from CEFR levels A2 to C2. The vertical axis shows the truncated PTE Academic theta scale varying from -2 to +2. The box plots show substantial overlap across adjacent CEFR categories, as well as an apparent ceiling effect at C2 for writing. CEFR levels, however, are not to be interpreted as mutually exclusive categories. Language development is continuous,

and does not take place in stages. Therefore, the CEFR scale and its levels should be interpreted as probabilistic: learners of a language are estimated most likely to be at a particular level, but this does not reduce to zero their probability to be at an adjacent level.

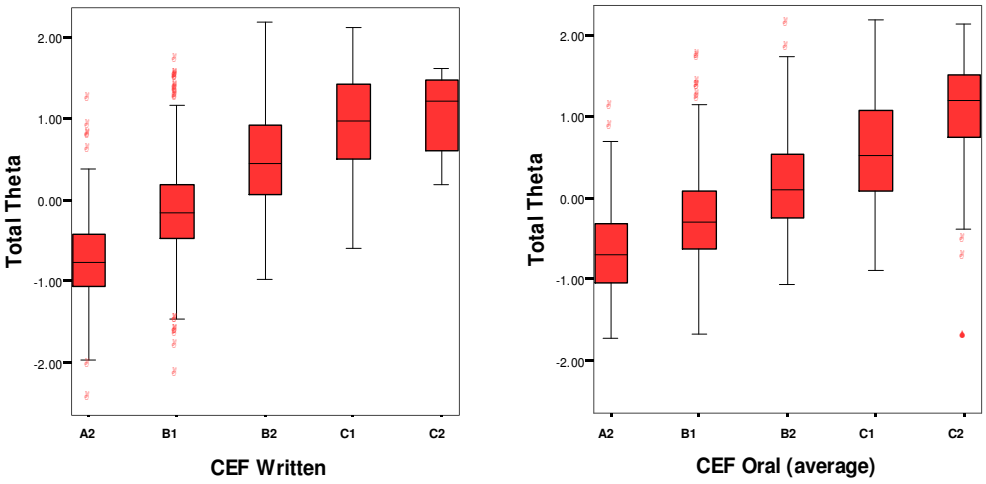


Figure 1: CEFR level distribution Box Plots

Although the official CEFR literature does not provide information on the minimum probability required to be at a CEFR level, the original scaling of the levels (North, 2000) is based on the Rasch model where cut-offs are defined at 0.5 probability. The distance of approximate 1 logit between levels implies that anyone typically reaching a probability of around 0.8 at level X, has 0.5 probability of being at level X+1 and is therefore exiting level X and entering level X+1. Having a probability of 0.5 of being at level X implies a probability of 0.15 to be at level X+1 and as little as 0.05 at level X+2. Based on the monotone increase of the PTE Academic theta from A2 to C2 as shown in Figure 2, a positive relation between the CEFR scale and the PTE Academic scale is established. To find the exact cut-offs on the PTE theta scale corresponding to the CEFR levels, the first stage is to establish the lower bounds of the CEFR categories based on the independent CEFR ratings. For this purpose, the CEFR ratings were scaled using FACETS (Linacre, 1988; 2005). The estimates of category boundaries on the CEFR theta scale are shown in Table 3.

Category	CEF Level	CEF Theta (Lower bounds)
0	BELOW A2	N/A
1	A2	-4.24
2	B1	-1.53
3	B2	0.63
4	C1	2.07
5	C2	3.07

Table 3: Category lower bounds on CEFR theta

The relationship between the scale underlying the CEFR levels and the PTE Academic theta for those test takers about whom we had information on both scales (n=3,318) is shown in Figure 2. The horizontal axis is the CEFR theta, and the vertical axis is the PTE

Academic theta estimate.The correlation between the two measures is 0.69. A better fitting regression is obtained with a first order polynomial (curved red line), yielding an r^2 of slightly over 0.5. This regression function was used to project the CEFR cut-offs from the CEFR scaled ratings onto the PTE Academic theta scale.

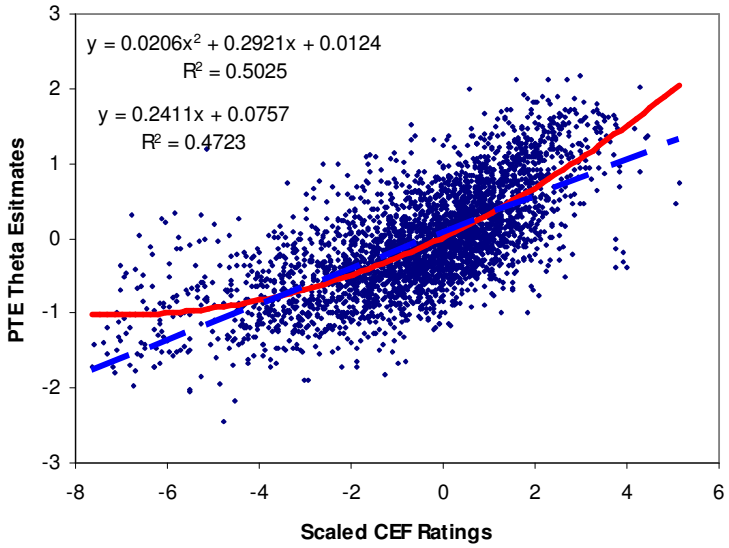


Figure 2: Relation between CEFR Theta and PTE Theta

Because of noisy data at the bottom end of the scales, the lowest performing 50 candidates were removed. Further analyses were conducted with the remaining 3,268 subjects. Figure 3 shows the cumulative frequencies for these 3,268 candidates for whom theta estimates are available on both scales (CEFR scale and PTE Academic scale). The cumulative frequencies are closely aligned though the PTE scale clearly shows smaller variance.

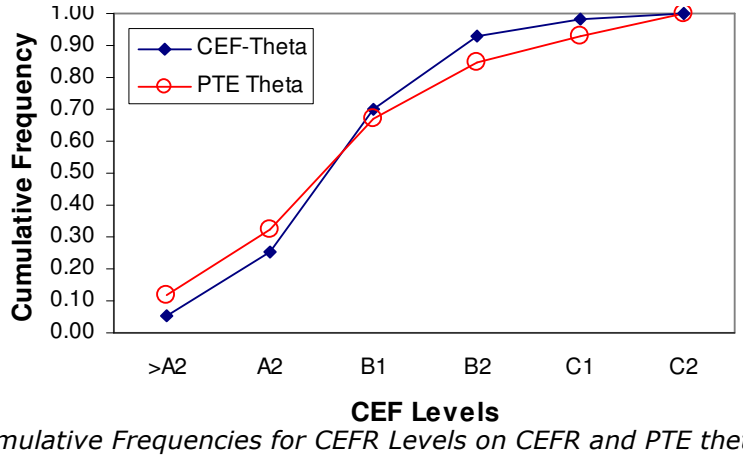


Figure 3: Cumulative Frequencies for CEFR Levels on CEFR and PTE theta scales

In the next stage, an equipercentile equating was chosen to express the CEFR lower bounds on the PTE theta scale. Equipercentile equating determines the equating relationship as one where a score has an equivalent percentile on either form. The cumulative frequencies are shown in Figure 4 and the projection of the CEFR lower bounds on the PTE theta scale together with the observed distribution of field test candidates over the CEFR levels is shown in Table 4.

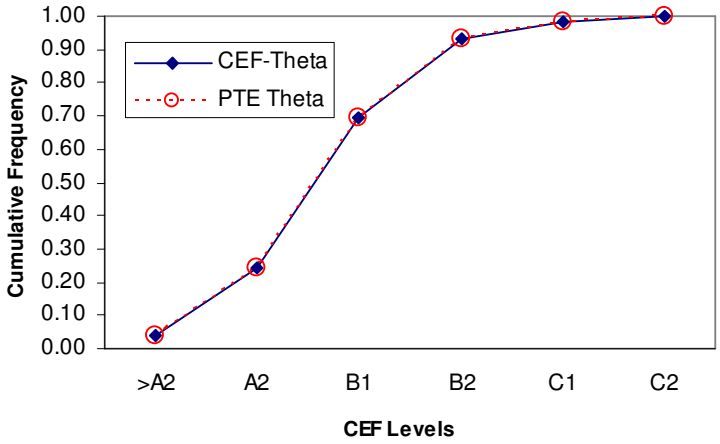


Figure 4: Cumulative frequencies on CEFR and PTE theta scales after equipercentile equating

CEFR Levels	Theta PTE	Frequency	Percentage	Cumulative Frequency
<A2	-1.366	126	4%	0.04
A2	-1.155	677	21%	0.25
B1	-0.496	1471	45%	0.70
B2	0.274	769	24%	0.93
C1	1.105	170	5%	0.98
C2	>1.554	55	2%	1.00
	Totals	3268	100%	

Table 4: Final Estimates for CEFR lower bounds on PTE theta scale

An Item-centered Approach

At the item development stage, item writers were required to indicate for each item which level of ability expressed in terms of the CEFR levels they intended to measure, i.e., did they think test takers would need to be able to correctly solve the items. In the item review process, these initial estimates from item writers were evaluated, and if needed, corrected by the item reviewers. Based on observations from field tests, the average item difficulty was calculated for items to fall into a particular category according to item writers. Table 5 provides the mean observed difficulty for each of the CEFR levels targeted by the item writers.

Intended CEFR Level	Mean observed difficulty
A2	0.172
B1	0.368
B2	0.823
C1	1.039
C2	1.323

Table 5: Intended and observed item difficulty

However, the cut-offs on the PTE Academic theta scale need to be established based on item writer estimates. To this effect, from the data, given item difficulty, the likelihood of any item to have been assigned to any of the CEFR levels was estimated. The cut-offs between the two consecutive levels is the location on the scale where the likelihood of belonging to the first category becomes less than the likelihood of belonging to the next category. In this way, the PTE theta cut-offs based on the items were found. The estimated lower bounds of the difficulty of items targeted at each of the CEFR levels were plotted against the lower bounds of these levels as estimated from the independent CEFR

ratings of test takers' responses by human raters. In Figure 5, the horizontal axis represents the CEFR cut-offs from the test taker-centered analysis, while the vertical axis represents the CEFR cut-offs from the item-centered analysis. Both estimates, derived independently, agree to a high degree (r=0.99).

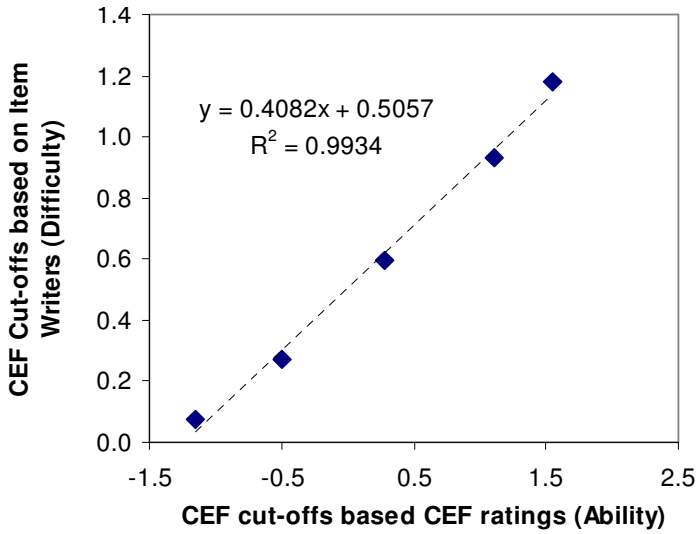


Figure 5: Lower bounds of CEFR levels based on targeted item difficulty versus lower bounds based on Equated CEFR ratings of candidates' responses

In addition, new test items are written by item writers adhering to both Test Specification and Item Writer Guidelines. These new items are systematically seeded in the operational test forms to gather live test taker responses. When enough responses are collected, these items are scored and analyzed together with other pre-calibrated test items. The analysis adopts a concurrent design of calibration, whereby new test items, following analysis of the results, are benchmarked to CEFR-referenced item difficulties.

Conclusions

Linking a test to a common scale like the CEFR presents its merits alongside its challenges. The processes of linking should involve intertwined stages from test development to statistical validation.

Considering the establishment of a solid link to the CEFR helps facilitate the interpretation of test scores to worldwide test users and potentially across tests of similar nature, this kind of establishment should be supported by both qualitative and quantitative evidence.

References

Linacre, J. M. (1988; 2005). A Computer Program for the Analysis of Multi-Faceted Data. Chicago, IL: Mesa Press.

North, B. (2000). The development of a common framework scale of language proficiency. New York: Peter Lang.

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. Measurement, 2, 178-182.

POSTER PRESENTATIONS

Samar Almoossa

Umm Alqura University, Makkah, Saudi Arabia

symossa@uqu.edu.sa

Are IELTS and CEFR Enough Indicator of Students Success in Academic Study?

Bio data

Samar Almoossa received her Master of Art in Teaching English as a Foreign Language from the University of Essex and is pursing her Doctorate of Philosophy in Applied Linguistics at Newcastle University. She works as a lecturer at the English Language center in Umm Alqura University. She has published three books in Arabic and has received awards from King Abdulaziz University in 2007 and 2008. In 2013, Almoossa participated in the UK Linguistics Olympiad (UKLO) as a marker, and she has been involved in postgraduate staff-student committees at Newcastle University.

Abstract

Language tests and the level achieved in the CEFR are considered to be the two main gatekeepers for international students who aim to study in an English-speaking university. However, gaining the required scores or passing from the A1 institutional level to the C1 level, instead of focusing on what they need to acquire for their academic study, has become the main aim for many students. Turner (2004, p.97) argues that "students seem to want to 'train' to reach the appropriate entrance level score or band rather than to engage with the language as an essential, and integral, part of the learning process of their subject of study."

The present study investigates students' perceptions of their experiences with language institutes, preparation for IELTS tests and how these two endeavors affected their academic study in UK universities. It also explores the relationship between the level these students achieved in the institutions according to the CEFR reference, the IELTS band score they attained and their academic performance. The study was based on a questionnaire that was distributed to 173 Saudi participants and on interviews undertaken with six of them.

The results revealed that participants' main concerns at every stage of their English language study was gaining the required score on the IELTS test, which led to giving more attention to the test and not preparing for their academic studies. Also, it was revealed that some of the participants in the study finished the advanced level (C1) in the institutions, yet they could not achieve higher than a score of 5-5.5 on the IELTS tests, which indicates the (B2) level in the CEFR. This study suggests that Practitioners should take IELTS band scores and CEFR "can do" statements with caution so students will clearly understand that what they achieved in term of scores or levels is not an indication of their readiness for academic study.

Short paper

Introduction

International students who are seeking to study in English-speaking universities must prove their readiness to join a program of study. Language tests are one of the main

official means used to evaluate student readiness. However, not all students can successfully attain the required scores on a language test, even if they have already studied all the levels in a language school. The present study aimed to examine the relationship between students' types of English, the levels achieved according to CEFR, their IELTS scores, and their satisfaction with their academic performance.

Method

The participants in the study were 173 Saudi Students from various disciplines in the UK. Most of them attended language institutions in the UK, studied various courses, achieved between A1 and C2 in terms of CEFR levels, and achieved between 4.5 and 8 on the IELTS test. The data were collected using a questionnaire that was administrated online. This was followed up with interviews of six participants in the study who had various experiences.

Findings

The results revealed that during their English language study, participants were oriented toward achieving certain CEFR levels and test scores, which caused a mismatch between their aims and the results achieved. Participants found that they were not well-prepared to respond to university academic demands. In addition, they reported that spending a great deal of time aiming to secure a given IELTS score and finish their language studies made the entire process more complicated and more difficult. Also, a huge mismatch was reported between the levels participants' achieved in language institutions, such as a C1 CEFR level, and IELTS band scores, which were often around 5-5.5. Such results suggest there is a need to view learners' achievements with caution. This study suggests that the alignment IELTS scores and CEFR levels could reveal a huge gap. This gap suggests that learners are not ready to study at English-speaking universities and that they should not be misled into thinking they are ready. Measurements such as IELTS and CEFR provide proof of learners' lingustic abilities, but they are not a substitute for developing general and academic language abilities. Practitioners should view IELTS band scores and CEFR "can do" statements with caustion so students will clearly understand that what they have achieved in term of scores or levels does not indicate their readiness for academic study. This study suggests that preparation for a language test that is meant to measure test takers' levels of proficiency and readiness to study at an English-speaking university should not mislead the learners and divert them from their main aim: learning the English language. The IELTS and CEFR, as measurements, provide proof of, but is not a substitute for, the development of general and academic language ability.

The link between CEFR specifications and learners' expectations

An understanding of how the European Common Framework (CEFR) contributes to identifying learners' linguistic ability levels is very important. The authors of this framework clearly declared, "We do not set out to tell practitioners what to do, or how to do it" (Council of Europe 2001, xi). There has been an ongoing effort since 1990 to align CEFR specifications with Cambridge ESOL tests in an attempt to gain a better understanding of test takers' band scores (Taylor, 2004). Khalifa and Ffrench (2009) declared that even language testers believe it is important to align the tests to the CEFR. On the other hand, test takers often question the link between their scores and the CEFR specifications (Taylor, 2004). CEFR "can do" statements provides learners with a guide to what they can achieve in their attempts to learn,. However, the findings of the current study suggest that learners who aim to purse an academic study, reach their CEFR levels and have good language test results should be guided these measurements are not indicators of readiness for academic study and success. Institutions and practitioners should make it clear to the learners that the CEFR specifications and test specifications are only guides. To conclude, the CEFR is strong tool to use in guiding learners, and practitioners should make it clear to learners how they can use its specifications to improve their language abilities.

References

Council of Europe. (2001). Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Khalifa, H. & Ffrench, A. (2009). Aligning Cambridge ESOL examinations to the CEFR: issues and practice. Cambridge ESOL: Research Notes 37: 10-14
URL: http://cambridgeesol.org/rs_notes/rs_nts37.pdf

Taylor, L. (2004). IELTS, Cambridge ESOL examinations and the Common European Framework. Research Notes, 18, 2-3.

Turner, J. (2004). Language as academic purpose. Journal of English for Academic Purposes, 3(2), 95-109.

Pilvi Alp, Krista Kerge & Hille Pajupuu

Institute of the Estonian Language, Tallinn, Estonia
Tallinn University, Tallinn, Estonia
Foundation Innove, Tallinn, Estonia

pilvi.alp@innove.ee - krista.kerge@tlu.ee - hille.pajupuu@eki.ee

Measuring Lexical Proficiency in L2 Creative Writing

Bio data

Pilvi Alp received her Masters degree in Estonian Philology from Tallinn University, currently studying for a PhD in Linguistics at Tallinn University. She works as a leader of Estonian language proficiency examinations development group at the Foundation Innove.

Fields of research: applied linguistics (linguistic markers of the different levels of Estonian language proficiency)

Current grants: Modelling and assessment of writing naturalness.
She has published in the areas of language testing and second language acquisition.

Kista Kerge (PhD) is Professor of Applied Linguistics at Tallinn University and a docent of language of law and administration at Tartu University, Estonia.

Fields of research: applied linguistics (linguistic standards and varieties; linguistic development in L1, L2); textlinguistics, genre analysis (mainly written text)
Current grants and projects: Modelling and assessment of writing naturalness (L1, L2); 3 projects on mother tongue teaching.

Publications: 10 monographs on Estonian syntax, functional literacy, B2/C1-level mastery of Estonian, word formation, and syntactic complicacy of texts; co-author of 8 books on L2-teachers' competency, B-level mastery, text-based language teaching, etc.; about 100 acad. articles.

Hille Pajupuu holds a PhD in linguistics. She is Senior Researcher at the Institute of the Estonian Language and a docent of intercultural communication at the Estonian Information Technology College, Estonia.

Fields of research: speech acoustics and perception; applied linguistics (language testing).

Current grants and projects: Modelling intermodular phenomena in Estonian; Statistical models of the emotionality of speech and written text; Modelling and assessment of writing naturalness.

Publications: 2 books in intercultural communication, 20 text books in applied linguistics (incl. as co-author), about 40 acad. papers.

Abstract

In order to learn how vocabulary changes with progress in language proficiency and which characteristics of the lexical proficiency should be taken into consideration when

assessing writing, we studied creative writings collected at official exams of Estonian as a second language (L2) on levels A2, B1, B2 and C1. For each level, we took for analysis 16 works from examinees who spoke Russian as their mother tongue and who had passed the writing assessment with a score of at least 70%. Vocabulary range was measured by comparing the words used with the frequency dictionary, resulting in a lexical frequency profile (LFP). Lexical diversity was measured via Guiraud's index (G) and the diversity of the sophisticated vocabulary via Advanced Guiraud (AG). The lexical sophistication (LS) and lexical density of the texts (LD) were calculated, as well. A two-sample t test was used to find the distinguishing characteristics of proficiency levels. For each level, all of the characteristics were correlated with the task scores. The outcome showed that LFP is similar as per level—that is, the writings consisted predominantly of frequent tokens and types. G was distinctive for all levels, increasing with language proficiency, but a positive correlation with score was established for B1 writings, only. AG was significantly different between A2/B1, A2/C1, B1/C1, B2/C1, respectively, but not between B2/B1 and B2/A2. AG correlated with scores only for A2 and C1 levels. LS differentiated between A2/C1, correlating also with the scores for these two. LD differentiated between A2/B1 and A2/B2, but did so without correlating with the scores. Assessors noted different lexical aspects in writings. Assessment guidelines must be complemented so that they take into account the features that distinguish between levels and so that they provide additional support for the assessors.

Short paper

Introduction

To bring the official job-related language proficiency exams into compliance with levels A2, B1, B2 and C1 as described in CEFR (2001), the extended-level descriptors of Estonian were accomplished (Hausenberg et al., 2008; Ilves, 2008, 2010; Kerge, 2008). Faced with a constricted time-frame for the implementation of the new Language Act (RT 2011), the preparation of tests and assessment guidelines was done mainly based on practitioners' experience. Recent analyses (e.g., Pajupuu et al., 2010; Türk et al., 2012) have highlighted the need to specify the selection of level feature aspects.

This survey is part of a wider research project that is aimed at identifying the characteristic features that enable the description and measuring of a learner's progress in writing skills. A major aspect of describing language proficiency levels is vocabulary, which is contained in both scaled descriptors and the assessment guidelines of CEFR. Vocabulary knowledge and use is of some import when writing in L2, with mastery of recurrent vocabulary being especially appreciated, while acknowledging the link between assessment of writings on a holistic scale and the author's knowledge of sophisticated words (Daller and Phelan, 2007; Milton, 2010; Stæhr, 2008; Yu, 2009). CEFR (2001: 5.2.1.1) describes lexical proficiency as vocabulary range and control. Judging by the descriptions offered by other researchers, an increase in communicative competence is related to command of a wider and more sophisticated vocabulary and more accurate word usage.

As some language-specific features of word-usage may not fully coincide with those presented in CEFR (cf., Milton and Alexiou, 2009), our goal is to elucidate how the lexical competence components—that is, vocabulary range, lexical diversity, sophistication and density of texts—are revealed in different-level Estonian writings, how they change with progress in the author's proficiency, how important they are when assessing different-level writings and how the requirements of vocabulary knowledge should be reflected in assessment guidelines.

Vocabulary assessment in local practice

In B1-, B2- and C1-level examinations of Estonian, three level-relevant aspects of written production are assessed: task completion, compositional organisation and linguistic range. The A2-level grading scale is comprised of two aspects—task completion and

linguistic range. According to the existing guidelines, each aspect must be scored separately; however, the final assessment reports only the aggregate score (the scores for each aspect added together). Each rating scale has been composed for a given task, the type and topic of which have been selected subject to the requirements of the level.
The rating scale has extended the description of level-specific lexical proficiency as part of linguistic range so that, in addition to excellent performance, moderate performance (that considered good and satisfactory) is reported on, thereby acknowledging the cut-off point of each level (Figure 1).

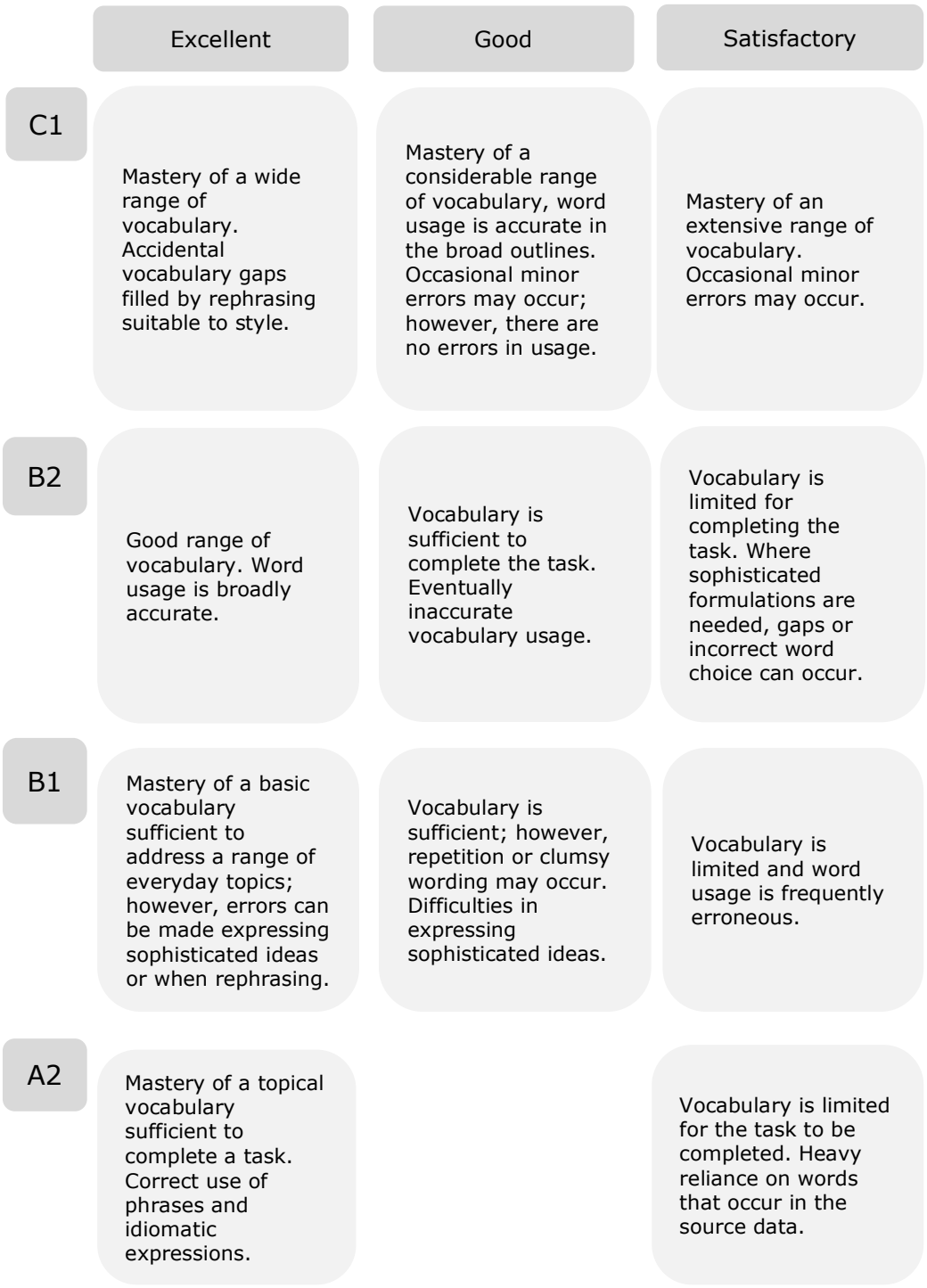


Figure 1. Synopsis of the lexical aspects of the rating scale.

The fact the description of lexical proficiency has no common denominator immediately catches the eye. In certain instances, lexical proficiency relates to a task – cf. A2 (excellent) and B2 (good): vocabulary is sufficient to complete the task; in other cases, it has been described in comparison with other levels – e.g., B1 (excellent): mastery of basic vocabulary, or B2 (excellent): good range of vocabulary.

Trained assessors orient themselves well in relation to the assessment guidelines of the respective levels; however, the synopsis for all levels (above) does not provide an overall view on the level-relevancy of vocabulary. The guidelines instead evidence the lack of precise knowledge as to how lexical proficiency is revealed in writings of different linguistic skill levels and what factors the assessor should heed in evaluating the author's vocabulary. Assessors clearly draw on prior assumptions in that, within the level, lexical proficiency can be assumed to be rateable on three levels: satisfactory, good and very good.

Background and research questions

It is expected that progress in the developing of writing skills is related to growth in vocabulary range and diversity, which enables the selecting of words and sentence structures that are more suitable for a given text-type or genre. In this regard, texts can be seen to become denser, more content rich and more nuanced in their lexical and stylistic choices.

In general, the learner first acquires the basic lexicon and thereafter acquires rare vocabulary (cf., Milton, 2010; Milton and Alexiou, 2009; Šišková, 2012); this position is also held by CEFR (2001: 112). Both the vocabulary range (cf., Van Hout and Vermeer, 2007) and the richness of infrequent vocabulary (cf., Nation, 2007) have been characterised by the text's lexical frequency profile (LFP), as described by Laufer and Nation (1995): the total vocabulary of the text is divided into frequency levels according to predetermined lists. The LFP shows the percentage of words (types or tokens) of different frequency-levels used in the text; the greater the number of words from the higher-frequency levels, the higher the proficiency level of the writer (Daller and Xue, 2007). It was found that the ratings were higher when the examinee used relatively fewer high-frequency tokens and relatively more low-frequency tokens. Furthermore, use of words that fall outside the list significantly increased the author's score (cf., East, 2009).

The lexical diversity of the text is a metric that illustrates how many unique words have been used in text. Diversity growth renders the messages more exact in a stepwise manner, producing a general picture of level-relevant vocabulary (cf., scales of lexical competence; CEFR 2001: 112). The most often-recurring technique used to measure lexical diversity is the type-token ratio (TTR). Although diversity growth indicates improvement in writing skills (e.g., Verspoor et al., 2012), its connection to rating is not as clear: assessors seem to be more influenced by sophisticated lexicons than by diversity (e.g., Daller and Phelan 2007; Van Hout and Vermeer, 2007). Verspoor et al. (2012) have noted the impact of diversity on scores for lower proficiency-level writings.

Sophistication of vocabulary is evidenced by the percentage of low-frequency words (advanced tokens) in a text (Laufer, 1995), which identifies the percentage of words in a given text that fall outside of what is considered to be a basic vocabulary. In English, this contains approximately 2,000 words covering some 80% of the text (Laufer and Nation, 1999). Language-wise, the number nevertheless may vary (Milton and Alexiou, 2009).

Rather than focusing on all types, Advanced Guiraud (Daller et al., 2003) is calculated by counting the number of advanced types per text and these are devided by the square root of the total number of tokens (cf. Tidball and Treffers-Daller, 2007).

Surveys have shown that the use of rare words is highly rated by assessors—that is, the lexical sophistication was broader in the higher-scoring essays. A tendency to focus on sophisticated words may be the result of an economical grading strategy (e.g., Daller and Phelan, 2007; East, 2009). The fact that users of rare words can be rated more highly, however, has also been noted by Nation (2007). The use of sophisticated words depends on text type, which need not be considered an invariably good discriminator with everyday topics within lower language levels (Verspoor et al., 2012).

Lexical density (Ure, 1971) is measured as a proportion of content words (e.g., nouns, verbs, adjectives and often adverb tokens, as well) to function words (e.g., prepositions, interjections, pronouns, conjunctions and count-word tokens). A text with a high proportion of content words gives more information (Johansson, 2008). Texts with higher density figures are likely to be richer and more demanding both linguistically and cognitively. The lower language proficiency levels, however, are not differentiated by density (Vidakovic and Barker, 2010). In such texts, density is affected by text type and the examinee's age (Johansson, 2008). It is not clear how density affects the scores.

Assuming that the vocabulary range, diversity, sophistication and text-density all exhibit growth as the examinee progresses from one proficiency level to the next, we have posed the following research questions for a comparative study of examination writings of authors operating at different proficiency levels:

1. Are the vocabulary range, diversity, sophistication and density sufficient to differentiate between the proficiency levels?
2. Is there a link between credits scored in writing task and the vocabulary range, diversity, sophistication and density of a text?
3. Should the assessors, by reference to results, focus separately on assessment of different aspects of lexical proficiency and, if so, when?

Material and method

The materials are derived from official, job-related, Estonian examinations. For papers of each level we analysed the pre-determined topic text written in a prescribed amount of time: A2-level, 30-word basic description; B1-level, 100-word detailed description (both 20 min); B2-level, 180-word verbal reasoning (50 min) and C1-level, 250-word publicist article (60 min).

Sixteen separate pieces of creative writing from each level were then subjected to analysis, with the sample limited to those papers that had successfully passed the level examination with a minimum score of 70%. The majority of writings analysed proved to be somewhat longer than was demanded (Table 1).

	A2			B1			**B2**			C1		
	N	**V**	**P**	N	V	P	**N**	**V**	**P**	N	V	P
max	85	50	6	149	93	10.5	301	136	10	378	183	12
Q3	50	36.2	5	129.2	75.5	8.9	275.2	118.5	9	335.8	158.8	11
median	40.5	29	4.5	122	71	8.25	221	101.5	8	269.5	139.5	10
Q1	35.5	25.2	4	101.7	59.2	7.5	190.5	96.7	8	258.8	135.5	9
min	31	23	4	81	48	6.5	160	84	7	236	111	9

Table 1. Tokens (N), types (V) and scores (P) per proficiency levels

Note. Maximum scores: level A2, 6 points; levels B1-C1, 12 points.

The native language of all examinees was Russian (age range 20–60 years, M=36.3, SD=11.6).

To measure vocabulary range, we compared the types and tokens of writings with the frequency dictionary of Estonian (Kaalep and Muischnek, 2002). We divided 10,000 words included in the dictionary into 10 separate 1,000-word tiers and calculated the LFP for each proficiency level.

General diversity was measured via Guiraud's index (Equation 1):

$$G=types / \sqrt{tokens} \quad (1)$$

where the greater the index, the more diversified the vocabulary.

The diversity of rare words was measured via Advanced Guiraud (Daller et al., 2003), which considered as advanced all word types – excluding names, numbers and abbreviations – that fell outside of the level of the 4,000 most frequently used words (Equation 2):

$$AG=advanced\ types / \sqrt{tokens} \quad (2)$$

The share of advanced tokens in a text was taken to be indicative of lexical sophistication (Laufer and Nation, 1995; Equation 3):

$$LS=advanced\ tokens *100 / total\ number\ of\ tokens \quad (3)$$

The share of content-word tokens in a text was taken to be indicative of lexical density (Ure, 1971); we considered as content words all nouns, adjectives, verbs and adverbs, with the exclusion of deictic adverbs (Puksand and Kerge, 2011; Equation 4).

$$LD=content\text{-}word\ tokens *100 / total\ number\ of\ tokens \quad (4)$$

A two-sample t test was used to identify the distinguishing characteristics of the proficiency levels. All characteristics were correlated with the task scores for each level.

Results

Vocabulary range

The LFP showed that the tokens of basic vocabulary in all writings covered approximately 90% (Figure 2) and types approximately 80% (Figure 3) of the text, while the share of rarer word types was relatively more—approximately 22%—for level C1.

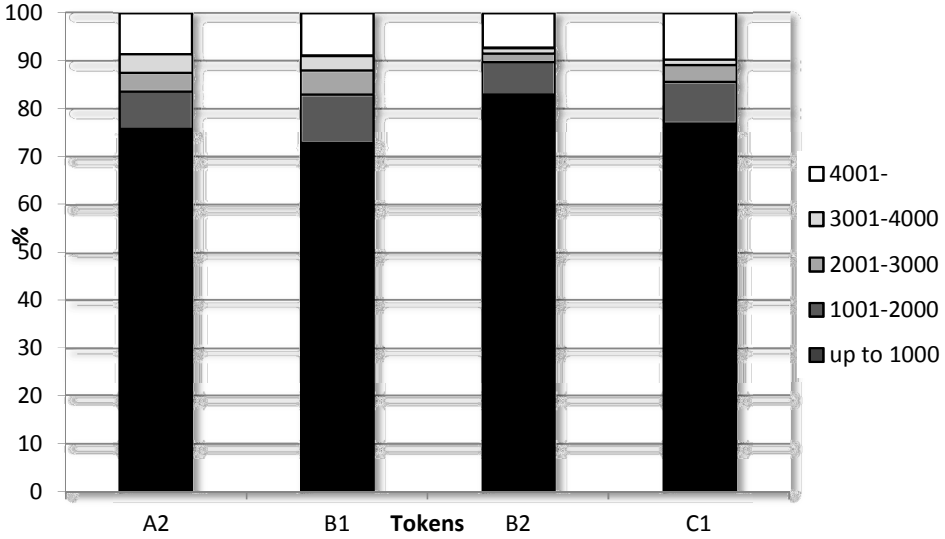


Figure 2. Share of tokens of different frequency in writings. Number of tokens: 738 (A1), 1,790 (B1), 3,495 (B2), 4,635 (C1).

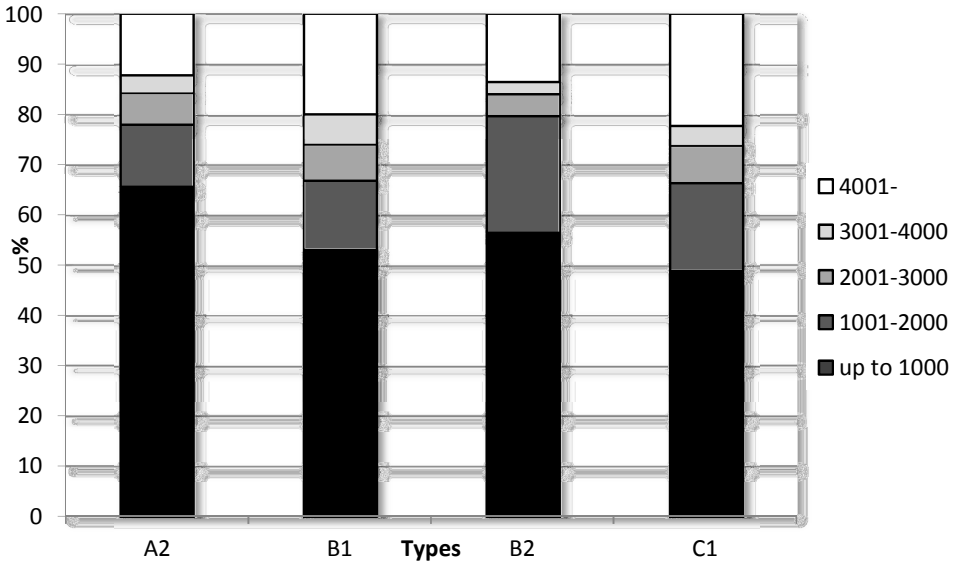


Figure 3. Share of types of different frequency in writings. Number of types: 289 (A1), 420 (B1), 679 (B2), 913 (C1).

Lexical diversity, sophistication and density

Average values of indices of vocabulary diversity, sophisticated-words diversity, lexical sophistication and density are presented in Table 2.

Indices	**G**				**AG**				**LS**				**LD**			
Levels	**A2**	**B1**	**B2**	**C1**	**A2**	**B1**	**B2**	**C1**	**A2**	**B1**	**B2**	**C1**	**A2**	**B1**	**B2**	**C1**
M	4.9	6.5	7.2	8.5	0.4	0.8	0.7	1.2	6.5	8.8	7.4	9.9	75.1	68.0	66.5	70.4
SD	0.5	0.8	0.6	0.9	0.2	0.4	0.2	0.3	2.3	4.2	3.0	2.4	7.3	5.0	3.6	5.3

Table 2. Indices of general diversity (G), diversity of sophisticated words (AG), lexical sophistication (LS) and density (LD)

Lexical diversity increased, keeping up with the progress of language proficiency and was and was a significant differentiator for all levels (Table 3; Figure 4).

Levels	**A2**	**B1**	**B2**
B1	.001***		
B2	.001***	.010**	
C1	.001***	.001***	.001***

Table 3. Two-sample t test for Guiraud's index

Note. *p<.05, **p<.01. ***p<.001.

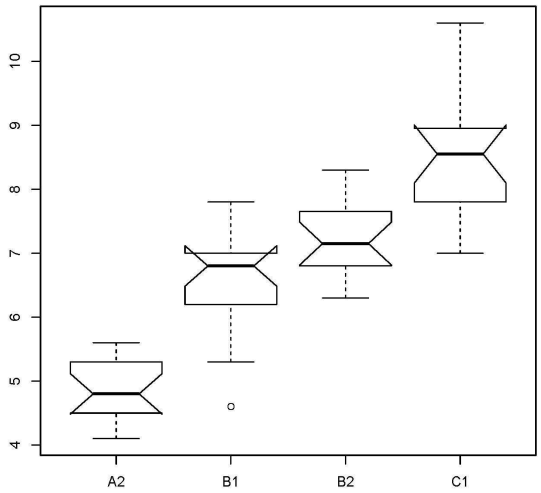


Figure 4. Guiraud's index.

The diversity of sophisticated word types revealed the tendency toward growth, again keeping with the progression of language proficiency. Significant differences were identified between levels A2/B1, A2/C1, B1/C1 and B2/C1; there was no significant difference noted between levels B1/B2 and B2/A2 (Table 4; Figure 5).

Levels	**A2**	**B1**	**B2**
B1	.001***		
B2	.053	.162	
C1	.001***	.001***	.001***

Table 4. Two-sample t test for Advanced Guiraud

Note. *p<.05, **p<.01. ***p<.001.

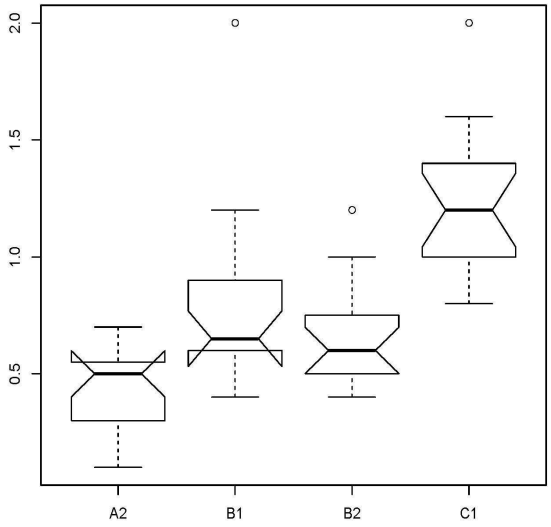


Figure 5. Advanced Guiraud.

Lexical sophistication also revealed the tendency for growth, in keeping with progress made in language proficiency; a significant difference was found only between levels A2 and C1 (Table 5; Figure 6).

Levels	A2	B1	B2
B1	.159		
B2	.705	.607	
C1	.021*	.705	.149

Table 5. Two-sample t test for lexical sophistication

Note. *p<.05, **p<.01. ***p<.001.

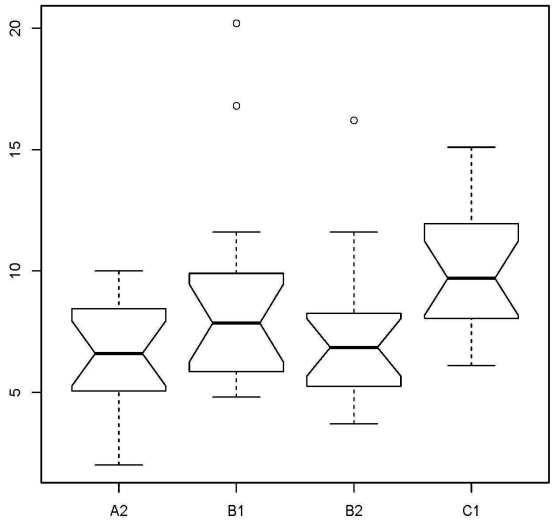


Figure 6. Lexical sophistication

Lexical density decreased, keeping with the progress of language proficiency. A significant difference was identified between levels A2/B1 and A2/B2. Between other levels, however, there were no significant differences noted (Table 6; Figure 7).

Levels	A2	B1	B2
B1	.004**		
B2	.001***	.481	
C1	.089	.481	.162

Table 6. Two-sample t test for lexical density

Note. *p<.05, **p<.01. ***p<.001.

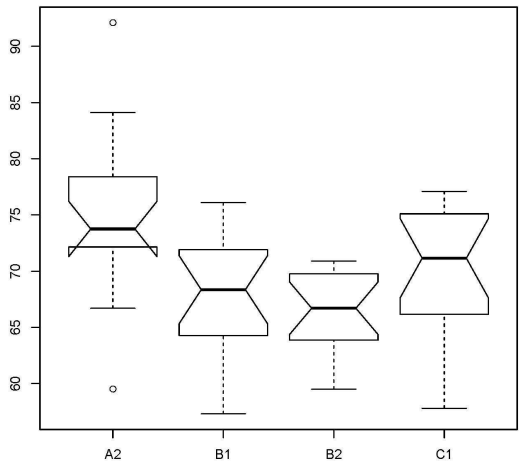


Figure 7. Lexical density

When correlating the vocabulary parameters with the scores for a given piece of writing, we noted that the vocabulary diversity (G) affected B1-level scores (the more diverse the vocabulary, the higher the score). At other levels, the diversity did not have any link with the reported scores. Indices related to vocabulary sophistication (AG and LS) were positively correlated with the scores in the case of A2- and C1-level writings. Lexical density was not linked to scores (Table 7).

	G	**AG**	**LS**	**LD**
C1	0.481	0.699**	0.736***	0.270
B2	-0.235	-0.016	0.299	0.025
B1	0.440*	0.357	0.208	0.480
A2	0.395	0.561*	0.584*	0.082

Table 7. Correlation between credit given for a written work and G, AG, LS and LD

Note. Two-tailed probability p has been marked as: *p<.05, **p<.01. ***p<.001.

Discussion

We studied the reflection of aspects related to the mastery of vocabulary in creative writings of authors at different language proficiency levels and its impact on scores, concentrating on vocabulary range, lexical diversity, sophisticated word diversity, lexical sophistication and density.

Both CEFR and guides for assessing the linguistic range of the writing test of the official Estonian examinations are based on the presumption that, with progress in language proficiency, vocabulary will supersede the framework of basic vocabulary and mastery of rare vocabulary will evolve (Daller et al., 2007; Milton, 2010; Stæhr, 2008; Šišková, 2012). We characterised the vocabulary range of each proficiency level with a LFP. Despite the data available to us, it was not possible to define a vocabulary range with reference to the writings because, regardless of their language-proficiency level, the examinees tended to rely heavily on basic vocabulary – for instance, frequent tokens accounted for approximately 90% and types for approximately 80% of the entire text. The same has been noted in texts of educated, native-Estonian speakers (e.g., Pajupuu et al., 2010).

Lexical diversity as measured via Guiraud's index revealed significant differentiation across all levels (e.g., Verspoor et al., 2012). Lexical diversity was shown to have increased with progress in language proficiency. It can be assumed that the assessors should intuitively pay attention to diversity; in this survey, however, the diversity and scores were positively related only at the B1-level. That partially bolsters the findings of Verspoor et al. (2012), in which the diversity of writings by authors with a low level of language proficiency as determined by Guiraud's index correlated with the holistically given scores. In the Estonian assessment guide, the importance of diversity has not been expressly pointed out, which may account for the assessors' apparent neglect of diversity. Due to the extensive scope and naturalness of frequent vocabulary, however, assessing general diversity may be found to be overly difficult.
Several analyses of general diversity and LFP have shown that sophisticated words used in writings have outsized impact on scores (e.g., Daller and Phelan, 2007; East, 2009; Van Hout and Vermeer, 2007). For Estonian in particular, the share of advanced lexis allowed for significant differentiation only between the levels A2 and C1. The most suitable vocabulary-assessing characteristic seems to be diversity of sophisticated words, which significantly differentiated all levels, excluding B1 and B2. As has been reported in prior surveys, the use and range of sophisticated vocabulary also affected Estonian assessors more than did any other vocabulary characteristic. Both share and diversity of sophisticated words were in positive correlation with the scores of writings at the A2- and C1-levels. While the assessment guide specifies a wide-ranging vocabulary as a criterion of the C1-level, the A2-level presumes knowledge of only elementary vocabulary (by frequency, the 2,000 most common words). Hence, dependence of the scores on the use of rare vocabulary as detected by assessors was not what we expected.

Lexical density significantly differentiated between the A2- and B1-levels and the A2- and B2-levels, respectively. The density of text decreased with progress made in language proficiency. Particularly dense A2-level texts were evidently a result of the shortness of the text (30 words) and the level-specific lack of natural means of coherence (e.g., discourse particles, variety of conjunctions, etc.), leading to texts with many simple sentences and few functional words. Scores were not affected by density.

The characteristics of the vocabulary of the writings can be used to differentiate between language-proficiency levels; however, assessors overlook these factors in favour of sophisticated words. The existing assessment guide offers little help in grading the vocabulary because it focuses – as per the example of CEFR – on vocabulary range in the first place, which does not significantly differentiate between writings of different levels. The assessment guide should seek to incorporate level-specific criteria for the measuring of lexical diversity as the primary distinguisher of levels. At the C1-level, it would be proper to draw the attention of assessors also to the diversity of sophisticated words.

There are some limitations to this work. We have not treated the issues related to the accuracy of word usage. It may prove an important vocabulary aspect for grading and is a topic that calls for further qualitative research. The present assessment guide contains some implicit suggestions for the appraisal of accuracy; however, it is not known how the assessors do or will view such guidelines. We also have not studied the changes of vocabulary characteristics within a level and therefore cannot answer the question of whether there is an indicator that would enable us to distinguish the weaker from the stronger examinee within the same level, or whether one description is sufficient to attribute a level of vocabulary to the writing without dividing it into satisfactory, good and very good.

To sum up, it is difficult to decide on vocabulary by reference to written tasks. It is, however, feasible, when the criteria of the lexical component in the assessment guidelines clearly define progress and provide support to assessor.

Acknowledgements
The study was supported by the Projects ETF8605 and SF0050023s09.

References

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. Journal of Second Language Writing, 18(3), 191-208.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press.

Daller, H., Milton, J. & Treffers-Daller, J. (eds) (2007). Modelling and Assessing Vocabulary Knowledge. Cambridge: CUP.

Daller, H., Van Hout, R. & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. Applied Linguistics, 24(2), 197-222.

Daller, H. & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, and J. Treffers-Daller (eds), Modelling and Assessing Vocabulary Knowledge (pp. 150-164). Cambridge: CUP.

Daller, H. & Phelan, D. (2007). Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Ed.), Modelling and assessing vocabulary knowledge (pp. 234-244). Cambridge: CUP.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. Assessing Writing, 14(2), 88-115.

Hausenberg, A.R., Kaivapalu, A., Kerge, K., Kern, K., Kitsnik, M., Krall, I., Rummo, K.& Rüütmaa, T. (2008). Iseseisev keelekasutaja. B1- ja B2-taseme eesti keele oskus. Tartu: Atlex.

Ilves, M. (2008). Algaja keelekasutaja: A2-taseme eesti keele oskus. Tallinn: Eesti Keele Sihtasutus.

Ilves, M. (2010). Läbimurre: A1-taseme eesti keele oskus. Tallinn: Eesti Keele Sihtasutus.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing : a develop- mental perspective. Working Papers, 53, 61-79.

Kaalep, H.-J. & Muischnek, K. (2002). Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli Kirjastus.

Kallas, J. & Tuulik, M. (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. ERÜ Aastaraamat, 7, 59-75.

Kerge, K. (2008). Vilunud keelekasutaja: C1-taseme eesti keele oskus. Tallinn: Eesti Keele Sihtasutus.

Laufer, B. (1995). Beyond 2000. A measure of productive lexicon in a second language. In L. Eubank, L. Selinker, & M. Sharwood Smith (Ed.), The current state of interlanguage (pp. 265-272). Amsterdam: John Benjamins.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. Applied Linguistics, 16, 307-322.

Laufer, B. & Nation, P. (1999). A vocabulary-size test of controlled productive ability. Language Testing, 16(1), 33–51.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Ed.). Communicative proficiency and language development: Intersections between SLA and language testing research. EUROSLA Monographs Series 1 (pp. , 211-232). European Second Language Association.

Milton, J. & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards et al. (Ed.), Vocabulary studies in first and second language acquisition (pp. 194-211). Basingstoke: Palgrave.

Nation, P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Ed.), Modelling and Assessing Vocabulary Knowledge (pp. 35-43). Cambridge: CUP.

Pajupuu, H., Kerge, K., Meister, L., Asu, E. L.& Alp, P. (2010). Natural speaking and how to assess it. Trames: Journal of the Humanities and Social Sciences, 59(2), 120-140.

Puksand, H & Kerge, K. (2012). Õpikuteksti analüüs kirjaoskuse omandamise kontekstis. Emakeele Seltsi aastaraamat, 2011(57), 162-217.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. Language Learning Journal, 36(2), 139-152.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. Language Studies Working Papers, 4, 26-36.

Tidball, F. & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. In H. Daller, J. Milton, & J. Treffers-Daller (Ed.), Modelling and assessing vocabulary knowledge (pp. 133-149). Cambridge: CUP.

RT 1995 = Elektrooniline Riigi Teataja. Kodakondsuse seadus. Vastu võetud 19.01.1995. Retrieved from https://www.riigiteataja.ee/akt/961169

RT 2007 = Elektrooniline Riigi Teataja. Keeleseaduse muutmise seadus. Vastu võetud 8. veebruaril 2007. Retrieved from https://www.riigiteataja.ee/akt/12795872.

Türk, Ü. & Kikerpill, T. (2012). Eksperthinnang eesti keele B2-taseme eksamitele 2009-2010. Analüüsi tulemused ja ettepanekud. Tartu: HTM.

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren, & J. L. M. Trim (Ed.). Applications of Linguistics. Selected Papers of the Second International Congress of Applied Linguistics, Cambridge 1969 (pp. 443-452). Cambridge: CUP.

van Hoult, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers Daller (eds). Modelling and Assessing Vocabulary Knowledge (pp. 93-115). Cambridge: CUP.

Verspoor, M., Schmid, M. S. & Xu, X. (2012). A dynamic usage based perspective on L2 writing. Journal of Second Language Writing, 21(3), 239-263.

Vidakovic, I. & Barker, F. (2010). Use of words and multi-word units in skills for life writing examinations. Research Notes, 41, 7-14.

Yu, G. (2009). Lexical Diversity in Writing and Speaking Task Performances. Applied Linguistics, 31(2), 236–259.

Maisa Martin, Ari Huhta & Riikka Alanen

University of Jyväskylä, Jyväskylä, Finland

maisa.martin@jyu.fi - ari.huhta@jyu.fi - riikka.a.alanen@jyu.fi

Using CEFR Scales in a SLA Study on Writing in a Second and Foreign Language

Bio data

Maisa Martin is the professor of Finnish as a second and foreign language at the Department of Languages. Her current research focuses on the development of functional writing skills in relation to the grammatical skills.

Ari Huhta is the professor of language assessment at the Centre for Applied Language Studies. He is widely experienced in a variety of forms of language testing and testing systems.

Riikka Alanen is the professor of foreign language pedagogy at the Department of Teacher Education. Her research interests revolve around the acquisition of Enlish in the Finnish context.

Abstract

The presentation is based on two projects that examine the linguistic (particularly grammatical and lexical) features and combinations of features that characterize different Common European Framework (CEFR) levels. The first project was a cross-sectional study of writing in English as a foreign language and Finnish as a second language in Finland (2007-09). The informants were 450 learners, aged 12 to 15, who completed four different functional writing tasks. The second project (2010-13) is a longitudinal study of the development of writing in English, Finnish and Swedish as L2 in Finland. A total of 550 learners, of different ages, complete the same or similar writing tasks each year for three years. The CEFR played a key role in the overall design of these studies as learners' writing performances were rated against the CEFR scales and against the more fine-tuned Finnish national curriculum scale (NC) based on the CEFR.

We present the results of multifaceted Rasch analyses on the CEFR and NC scales to show how they worked for rating purposes; this is particularly interesting as there are few published studies on the use of unmodified CEFR scales for rating purposes. Research into the quality of the Finnish NC scale is almost completely lacking.

As an example of a qualitative analysis of the two scales we present the findings based on 4300 comments written on the rating forms by the raters when assessing the learners' texts. We focus on the use of the term 'simple' in the CEFR scales and compare the raters' interpretations of the term with the contexts it is used in the CEFR scales to determine what the dimensions of simple in the CEFR are and in what way they are connected to the issues in the complexity discussion within the study of second language development.

Short paper

Second language development (SLD) involves both functional and lexical-structural growth. The former is the main focus of CEFR as the aims and purposes of language use are much the same across Europe and European languages. Structural and lexical issues are referred to and some very general scales are provided but as the ways the lexical and even more the structural skills develop in widely varying ways across languages, it is necessary for practical purposes to study the linguistic development also at the level of individual languages. At the same time such studies provide either supporting or countering evidence for SLD theories which claim to be universal.

This paper and the accompanying poster are based on two projects that examine the linguistic (particularly grammatical and lexical) features and combinations of features that characterize different Common European Framework (CEFR) levels. The first project is a cross-sectional study of writing in English as a foreign language and Finnish as a second language in Finland (see Cefling 2007-09). The informants were 450 learners, aged 12 to 15, who completed four different functional writing tasks. The second project (see Topling 2010-13) is a longitudinal study of the development of writing in English, Finnish and Swedish as L2 in Finland. A total of 550 learners, of different ages, completed the same or similar writing tasks each year for three years.

The main aim of the projects is to study the structural and lexical development at the CEFR levels A1–C2 across languages and age groups. The CEFR played a key role in the overall design of these studies as learners' writing performances were rated against the CEFR scales and against the more fine-tuned Finnish national core curriculum scale (NCC) based on the CEFR. Another aim is to compare the two scales and evaluate them as rating tools. Rater behavior is also in focus, approached through rater interviews and the study of the comments raters wrote on rating forms. The results presented here are mainly based on the cross-sectional Cefling project, as few developmental results of the longitudinal Topling project are available at the moment. The results on the rater comments include some Topling data as well.

In their different phases these studies have addressed more or less all the themes of the conference, with the main focus on the first one. In this paper we focus on the assumptions which underlie the efforts to create links between the functional CEFR levels and language-specific or general SLD. To be able to provide such links, CEFR levels, or levels of other similar scales as NCC in our case, must be reliably established. The results of this part of the project confirm the validity of the linguistic trends found in the other studies of these projects (for publications and presentations, see the projects' webpages).

Neither of the scales was originally designed for rating purposes although both are commonly used as such, and it is interesting that there appears to be little research on their suitability for rating (certainly none for the NCC scales). Here the two scales were examined to see how well they worked for rating purposes. For this aim, multifaceted Rasch programme Facets analyses were conducted, according to the guidelines presented in previous studies (Linacre, 2009; Lunz, Wright & Linacre, 1990; McNamara, 1996).

Factors such as task difficulty and the varying leniency or strictness of the raters obviously affect the results. These factors are incorporated in the Facets analysis. As the rating design involved several raters for each text, it was possible to remove one inconsistent rater and occasional inconsistent ratings from the data. The iterative analyses indicate that the quality of the ratings of the scripts in the final project data set was consistent enough that the placements on the CEFR and NCC levels can be trusted.

A good rating scale should also separate the proficiency levels from each other. Figure 1 shows the shape of the CEFR scale used for the rating of English writing performances

based on all ratings across all tasks and raters. The four 'hilltops' represent, from left to right, levels A1, A2, B1 and B2; the slope left of A1 stands for 'under A1' and the slope right of B2 represents C1. We can immediately see that the levels are in the correct order and that they are well separated from each other. Linacre (2002) argues that the minimum distance between scale points should be 1.4 logits for them to be clearly separable; this is however related to scale length so that for a 5-point scale 1.0 logit separation should suffice. This criterion is clearly met for the CEFR scale. The same is true when the scales are applied to Finnish as a second language.

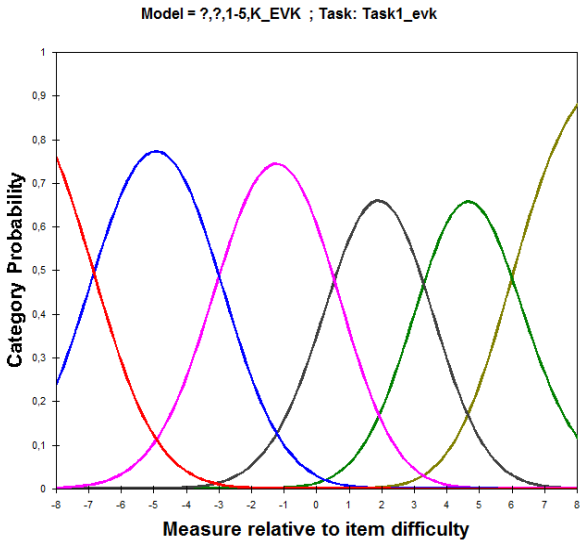


Figure 1: Shape of the CEFR scale in the English rating data

Figure 2 displays the shape of the National Curriculum scale. Again, the levels run from A1.1 on the left to B2.2 on the right, with the extreme ends (slopes) representing 'under A1.1' and C1.1. There are no disordered levels and each 'hilltop' can be distinguished, although some levels are rather narrow suggesting that raters may have some problems with them. The narrowest levels for English are A1.1 and B1.1. However, given the length of the NCC scale, even these levels are probably not problematic. Again the Finnish data yields similar results. (For more on the technical background and details about the figures, see Huhta et al. 2013.)

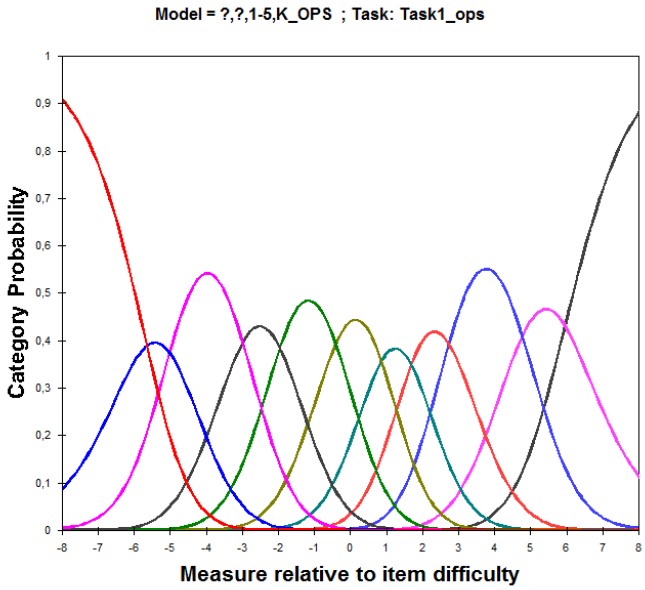


Figure 2: Shape of the National Curriculum scale in the English rating data

In overall comparison the two scales meet quite well: the cut-off point between e.g. A2.2. and B.1.1. in the ten-level NCC scales is very close to that of A2 and B1 in the CEFR scales. If anything, the NCC scales are slightly more demanding, particularly for English the NCC boundaries were consistently somewhat higher than the CEFR boundaries. A borderline A1/A2 learner was likely to be rated as A2 on the CEFR scale but A1.3 on the NCC scale. This may be at least partly due to the frequent reference to errors in NCC, which was not present in the CEFR scales used in this study.

In addition to the extensive statistical analysis, of which only some results are presented above, the scales were also subjected to a qualitative analysis. The CEFR scales have been criticized for ambiguities and inconsistencies and for lack of definitions (see e.g. Alderson, 2007, 661). The NCC scales are often discussed at teachers' events but no formal studies of them exist before the current one.

The raters in these projects worked with forms which allowed them to write one main rating, a seondary rating if they felt it necessary, and an optional comment. The 4300 comments were collected, typed, and classified by a group of students[1] who also helped to analyze the data. The comments nearly always refer to some problem in rating (no one wrote "This is easy!" or "I am sure of this."), whether specific to the script to be assesssed or issues relating to the rating scales, or both. Certain comments were personal in nature, e.g. some raters preferring the six-level scale while others found it easier to work with the ten-level scale. But mainly the comments, as well as the rater interviews also conducted as part of the projects, revealed a number of dubious wordings in the scales. Here we focus on only one of them: the use of the term 'simple' in the CEFR scales.

Second language development is often described using three trajectories: complexity, accuracy, and fluency (the CAF triad, see e.g. Housen et al. 2012). Learners are assumed to move from simple to complex, as is testified also by the CEFR scales where the word

[1] We thank Johanna Eloranta, Milla Filppula, Marjaana Göös, Raisa Haikala, Heidi Henttinen, Ulla Huhtala, Janica Häggman, Mari Karppinen, Milja Koski-Lammi, Auli Kotimäki, Maiju Partanen and Elisa Räsänen.

simple appears from A1 to B1 and the word complex at the levels C1 and C2. The level B2 seems to be neither simple nor complex. The level A2 is particularly simple: the word appears in the writing descriptors[2] 13 times!

Complexity in CAF research is most commonly determined by counting some linguistic features, assumed to be complex for learners, in learner production, either speech or writing. Typical complexity indices include the number of subordinations per sentence or morphemes per clause etc. (for overviews see e.g. Wolfe-Quintero et al. 1998, Ortega 2003). An example of a more qualitative, distribution-based approach to complexity is Reiman 2011. None of these, however, seem to have much to do with simple or complex in the CEFR.

What collocates with simple in CEFR? Most commonly simple refers to genres: simple postcard, note, list, letter, message, poem, biography. It appears also with grammatical terms: simple phrase, sentence, element, connector, text. Content is rarely referred to in CEFR, but complex is connected with the subject of the text at the level C1. In addition, words which could be interpereted as (near) synonyms os simple, such as basic or straightforward, also appear in the scales. With such a variety of usage and counterparts in the descriptors, it is no wonder that simple is hard for the raters to interpret. This is reflected in the comments. Many raters ponder, whether simple means the same as short. In particular in the contexts where simple is attached to a genre, raters connect it to limited content. Yet another, quite common, context in the rater comments about simple is linguistic: limited syntax or vocabulary.

The linguistic interpretation of simple leads to further questions. As measures of syntactic complexity in SLD studies abound, which one(s) are we talking about? Should raters pay attention to them, at least as some rough evaluation? Or, if the development from simple to complex is crucial at some levels (as the number of references to it at A2 seems to indicate), should we aim at developing programmes which would automatically rate writing samples based on some complexity measures? Some such programmes already exist, e.g. Cohmetrix for English or Direkt Profil for French). How about lexical complexity? Many automatic analyzers for it can be found, too, but all of these share two weaknesses: They do not exist for all languages and usually require the texts to be standardized as they cannot handle linguistic errors - which of course are common in learner language particularly at the lowest levels.

The brief list of questions above, all around one word in the CEFR descriptors, opens a can of worms. A set of scales intended to be used across all languages and all levels of proficiency, in many different learning and teaching contexts, can never be perfect. Nevertheless, a closer co-operation with those involved in SLD studies, particularly the CAF triad, might be useful in developing the scales into a more systematic direction, involving references to complexity, accuracy and fluency at all levels, not just here and there. Furthermore, linguistic complexity should be separated from the complexity of content and genre by more careful choices of terms and better definitions of them.

Even if a detailed analysis of the CEFR descriptors, such as about the word simple above, can easily reveal gaps in the CEFR thinking, the results presented in the first part of this article are encouraging. With the help of the statistical analysis taking into account a variety of factors influencing the assessment process, a fairly reliable assessment can be achieved, at least where the raters share the first language, the same cultural and educational background, and are experienced and repeatedly trained before each rating session. This is what matters most for the learners and educational institutions. But it also leaves room for SLD researchers, in cooperation with language testers to examine the CEFR scales to improve them to better match what is known of the parameters of SLD and to be more systematic in content and presentation.

[2] Only the functional scales were used in these studies, excluding the ones referring to linguistic profieciency and errors.

References

Alderson, J.C. (2007). The CEFR and the Need for More Research. The Modern Language Journal 91(4), 659-663.

Cefling. (2007-2009). https://www.jyu.fi/cefling

Cohmetrix: http://cohmetrix.memphis.edu/

Direkt profil: http://dl.acm.org/citation.cfm?id=1609838

Housen, A., Kuiken F. & Vedder, I. (2012). Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA. Amsterdam: John Benjamins.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M. & Hirvelä, T. (2013): Assessing learners' writing skills in a SLA study - Validating the rating process across tasks, scales and languages. Manuscript submitted for publication.

Linacre, M. (2002). Optimizing rating scale category effectiveness. Journal of Applied Measurement 3, 2002, 85-106.

Linacre, M. (2009). A User's Guide to FACETS v 3.66.0. Chicago: Winsteps.

Lunz, M., Wright, B. & Linacre, M. (1990). Measuring the impact of judge severity on examination scores. Applied Measurement in Education, 3(4), 331-345.

McNamara, T. (1996). Measuring Second Language Performance. Boston: Addison-Wesley Longman.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics, 24(4), 492–518.

Reiman, N. (2011). Two faces of complexity: structural measures and diversity of constructions. Nordand 6(2), 9-33.

Topling. (2010–2013). https://www.jyu.fi/topling

Wolfe-Quintero, K., Inagaki, S. & Kim, H.-E. (1998). Second language development in writing: measures of fluency, accuracy, and complexity. Technical report #17. Second Language Teaching & Curriculum Center. University of Hawai'i at Mãnoa.

Institute of Education and Information Sciences - Linguapolis

University of Antwerp | Prinsstraat 13 | 2000 Antwerp | Belgium

www.ua.ac.be/lt-cefr2013

View publication stats