An LSP Learner Corpus to Help with English Radiotelephony Teaching

Stéphanie Lopez, Anne Condamines, Amélie Josselin-Leray

HAL Id: halshs-01380802
https://shs.hal.science/halshs-01380802

Submitted on 13 Oct 2016

An LSP Learner Corpus to Help with English Radiotelephony Teaching

Stéphanie Lopez
CLLE-ERSS, CNRS & Université de Toulouse
& ENAC[1]

Anne Condamines
CLLE-ERSS, CNRS & Université de Toulouse

Amélie Josselin-Leray
CLLE-ERSS, CNRS & Université de Toulouse

Abstract

The French Civil Aviation University (ENAC) is in charge of the French controllers' initial training in English and has therefore specific needs in terms of English radiotelephony teaching. Consequently, an observation of the usage of English made by French controllers with international pilots, that is to say ongoing foreign language learners, was initiated. The aim of this project is to describe and categorise the different uses of English within pilot-controller communications through the means of a comparative study between two corpora. The ultimate purpose of this comparative analysis is foreign language (English for Specific Purposes) teaching.

Keywords: Language for Specific Purposes, English radiotelephony communications, modal verbs.

1.   Introduction

Over the past 20 years, many contributions have been made to the field of Learner Corpus Research (see among others Tono 2000; Granger *et al.* 2002; Nesselhauf 2004; Gilquin *et al.* 2008; Myles 2008; Mukherjee 2009; Granger & Gilquin 2011). Nevertheless, within this relatively recent field of research, there has been little interest in Language for Specific Purposes (LSP) Learner Corpora (Granger & Paquot forthcoming). Moreover, the difficult and time-consuming task of analysing spoken data in general makes *spoken* LSP learner corpora very rare. The corpus at the core of our research project is one of the few that deal with spoken LSP and in particular with controlled languages. It is not, however, a typical learner corpus, insofar as it deals with the language produced by English language learners *outside* an educational setting (*ibid.*). Nevertheless, since the speakers of our corpus are subject to ongoing foreign language learning throughout their careers and since the ultimate purpose of our corpus is foreign language (English for Specific Purposes) teaching, we like to think of it as an original learner corpus. It consists of the controlled language used in a very specific domain: that of air traffic control.

---

[1] École Nationale de l'Aviation Civile.

In this paper, we aim to show that the use of an LSP advanced learner corpus can be a very useful pedagogical tool. The observation of regular usage patterns, whether correct or incorrect, could indeed help in compiling up-to-date materials to meet the (current and future) controllers' needs and interests. First, the notions of English phraseology and plain language, the two types of languages used in pilot-controller communications, are introduced. Then, we present the two corpora under study. Finally, we focus, as an example, on the way English modals are used – particularly how the modal *can* is used by French controllers – before giving some perspectives on possible pedagogical applications of our results.

2. Context of the Study

Pilot-controller communications are mainly performed by means of a Language for Specific Purposes called *phraseology*[2]. It was created to try and ensure safety in the most common radiotelephony communications. In situations for which phraseology has not been set, pilots and controllers must use *plain language*. In the case of international flights, English has to be available for flight crews who do not speak the language of the control centre on the ground. In French controlled airspace, for instance, French controllers are able to deal with pilots from all over the world thanks to English phraseology and plain language. Training and evaluation of their English throughout their entire career is compulsory to ensure the safest possible communications. French controllers can therefore be considered to be permanent learners of English at an advanced level since they possess a certain operational level in English.

The ENAC – the French Civil Aviation University – is in charge of the French controllers' initial training in English and has to comply with internationally prescribed language proficiency requirements. It has specific needs in terms of English radiotelephony teaching. Consequently, it has initiated, in collaboration with the research institute CLLE-ERSS, an observation of the usage of English made by French controllers (Lopez *et al.* 2011). The aim of this project is to describe and categorise the different uses of English phraseology and plain language within air-ground communications thanks to a comparative study between two corpora. One of these corpora – the LSP learner corpus – and the results from the analysis can be used for English teaching purposes such as developing appropriate and up-to-date pedagogical materials.

3. English Phraseology and Plain Language

Phraseology, as adumbrated above, has been designed to meet specific communication needs: it was created and has been continually updated to cover the most common and ordinary situations encountered in air navigation. Its aim is to ensure non-ambiguous and effective communications between pilots and controllers through simplified

---

[2] The term introduced here has therefore nothing to do with the notion of phraseology commonly used in linguistics.

syntactic, lexical and semantic rules: "the purpose of phraseologies is to provide clear, concise, unambiguous language to communicate messages of a routine nature" (ICAO 2010 [2004]: 1.1.3). Its main characteristics can be summed up as (Mell 1992; Philps 1989, 1991; Rubenbauer 2009) the omnipresence of the imperative form in controllers' messages; the rarity of the interrogative and negative forms; the almost complete absence of modals; the deletion of determiners, subject pronouns and prepositions; the deletion of auxiliaries *be* and *have* in *be + past participle* forms, *be + -ing* forms and *have + past participle* forms; the nominalisation of concepts; a highly specialised, univocal and finite lexicon; the use of an alphabet proper to the aeronautical domain; and finally the specific spelling and pronunciation of numbers. This type of language for specific purposes being seldom encountered, the following example provides an idea of what phraseology is like:[3]

P: Mérignac Delivery, Citron Air 3 2 4 5, stand delta 8, requesting start-up, destination Lyon, information lima.
C: Citron Air 3 2 4 5, start-up approved, CTOT 1 2 5 0, Sauveterre 3 alpha departure, level 1 1 0, squawk 2 3 2 0.
P: roger, start-up approved, CTOT 1 2 5 0, Sauveterre 3 alpha, level 1 1 0, squawk 2 3 2 0, Citron Air 3 2 4 5.

The syntactic, lexical and semantic characteristics of phraseology make it the essential communication tool for the transmission of the specialised knowledge required by pilots and controllers to do their jobs, and provide optimal and safe air traffic control. Phraseology is, however, a limited tool since it was created to cover only a limited number of air navigation situations. As a consequence, when facing situations for which phraseology does not exist, pilots and controllers must resort to *plain language*. We should specify here that the notion of plain language within the domain of air traffic control has a definition of its own:

> Plain language in aeronautical radiotelephony communications means the spontaneous, creative and non-coded use of a given natural language, although constrained by the functions and topics (aviation and non-aviation) that are required by aeronautical radiotelephony communications, as well as by specific safety-critical requirements for intelligibility, directness, appropriacy, non-ambiguity and concision (ICAO 2010 [2004]: 3.3.14).

The notion of plain language, despite its recent formal definition, is not always clearly understood by professionals in charge of English radiotelephony teaching: how can a spontaneous, creative and non-coded language reflect the features of clarity, concision and preciseness that must characterise radiotelephony communications? In order to have a better idea of the actual use of English plain language by French controllers, alongside the use of English phraseology, a comparative study between two corpora is being carried out with the ultimate goal of contributing to English radiotelephony teaching. Such appropriate corpora did not exist and had to be compiled.

4. Corpora under Study

---

[3] Messages starting with *P:* correspond to pilots' messages; the ones starting with *C:* correspond to controllers'.

One of the methodologies commonly used in learner corpus research in order to try and identify second language specific features is known as *contrastive interlanguage analysis* (Gilquin *et al.* 2007). One branch of this type of analysis consists in comparing learner with native or expert language (*ibid.*). In our case, since no one can really be considered as a native speaker of English radiotelephony – phraseology is obviously a designed controlled language presenting characteristics that are far too specific for anyone to acquire it as their first language – we have resolved to base our contrastive interlanguage analysis on a corpus made up of standardised official phraseology (henceforth referred to as *reference corpus*), and a corpus representing the language used by ongoing English phraseology and plain language (*i.e.* English radiotelephony) advanced learners, that is to say, French controllers and international pilots (henceforth referred to as *LSP learner corpus*).

The first step in the compilation of our reference corpus was to select official texts from which representative samples of standard phraseology could be extracted. This type of texts being quite rare, all the examples in English from two phraseology manuals – one edited by the ICAO (2007 [1990]) and the other by the French government (DGAC 2007 [2000]) – have been selected. By choosing these two different manuals, we aim at representing the norm from an international as well as a national point of view.

Our LSP learner corpus consists of the orthographic transcriptions of about 22 hours of recordings of real pilot-controller communications in English. The recordings were first collected in three different French control centres[4] in order to try and ensure that the corpus is representative of the language used in different types of air navigation control situations. A specific transcription protocol has then been created and the transcriptions obtained have been reviewed by air traffic control experts.

The quantitative characteristics of the corpora under investigation, in terms of tokens and types, are presented in the following table:

	Reference corpus			LSP learner corpus			
	Manual 1	Manual 2	Total	Centre 1	Centre 2	Centre 3	Total
Tokens	9,112	7,169	16,281	24,458	16,969	35,755	77,182
Types	659	656	867	731	564	836	1,282

Table 1. Size of the Corpora

One can notice that the number of word *types* contained in the corpora remains rather low despite the total number of *tokens*: it results from the finiteness of the phraseology lexicon mentioned earlier. Furthermore, the total number of word types found in the corpora is not equal to the sum of word types included in each of their respective sub-corpus because the latter share several common word types; *e.g.*, the word types "will" and "should" are found in both manuals.

---

[4] To collect these communications, an official authorisation was needed beforehand as, in France (and many other countries), this type of data is not accessible to the general public. For reasons of anonymity, the names and locations of the three centres are not revealed in this paper. They have been chosen for the concentration of English used on their frequencies as well as for their interest in our research project.

Although our corpora can be considered as small, we have chosen to reduce further the number of lexical forms on which to base our analysis by cleaning them up. Our purpose by doing so is to try and draw a well-balanced comparison of the lexicon found in the corpora. Comparing, from a lexical point of view, a corpus made up of written data – therefore including none of the features of verbal communication – with one made up of spoken data would not guarantee satisfactory results. Thus, words in languages other than English, truncated words and *hapax legomena* (words with only one occurrence) found in the LSP learner corpus, as well as proper nouns and numbers[5] from both corpora are not taken into account in our comparative analysis. By excluding these word forms, we aim at a well-proportioned comparative study focussing on specific and recurrent air traffic control vocabulary. The two corpora revised for the goals of our study thus contain fewer word types and tokens: the reference corpus comprises 7,181 tokens of 670 word types, while the LSP learner corpus contains 24,313 tokens of 495 types.

5.   Comparing the Corpora

Owing to their particular syntactic structures, neither of the corpora could be subjected to a correct computer morpho-syntactic tagging. The preliminary step in comparing the lexicon of the two corpora was to manually classify all the word types they comprise according to their part-of-speech or grammatical categories. The classification displays several salient discrepancies between the reference corpus and the LSP learner corpus, especially among the noun, verb, adjective, interjection, determiners and pronoun categories, as illustrated in Figure 1. For instance, the distribution of the categories shows that, in our context of study, English radiotelephony learners tend to use nouns to a lesser extent than advocated by the manuals (34.8% of all tokens in the LSP learner corpus *vs*. 47.4% in the reference corpus). The presence of twice as many noun types in the reference corpus (302 noun types) as in the LSP learner corpus (148 noun types) can of course be explained by the fact that the former is constituted of examples from manuals which try to be as exhaustive as possible, referring to various types of air navigation situations, some of which are not represented in the LSP learner corpus. The subcategory of acronyms is a good example of this conjecture: it represents 8.2% of all the noun tokens from the reference corpus, with 25 domain-specific types of acronyms such as QNH[6], ILS (Instrument Landing System), VFR (Visual Flight Rules), IFR (Instrument Flight Rules), NDB (Non-Directional Beacon), RVSM (Reduced Vertical Separation Minima), CTOT (Calculated Take-Off Time), CBs (Cumulonimbus – a type of cloud) or TCAS (Traffic Collision Avoidance System). Conversely, in the LSP learner corpus, acronyms represent only 3.1% of all the noun tokens, with only 3 types of acronyms: QNH, ILS and CBs. But the observed underuse of nouns by English radiotelephony learners may also result from a non-compliance with the nominalisation process which particularly characterises phraseology. In any case, this underuse of nouns can account, to all appearances, for the learners' overuse of verbs.

---

[5] There is a large proportion of numerical references in both corpora. They represent 35,9% of the total number of tokens from the reference and 43,72% of the LSP learner corpus

6 "QNH" is not an acronym *per se*: it is one of the few "Q codes" still in use to refer to a specific pressure setting which can be defined as the barometric pressure adjusted to sea level.

Figure 1. Distribution of the lexical units in the corpora according to their parts of speech
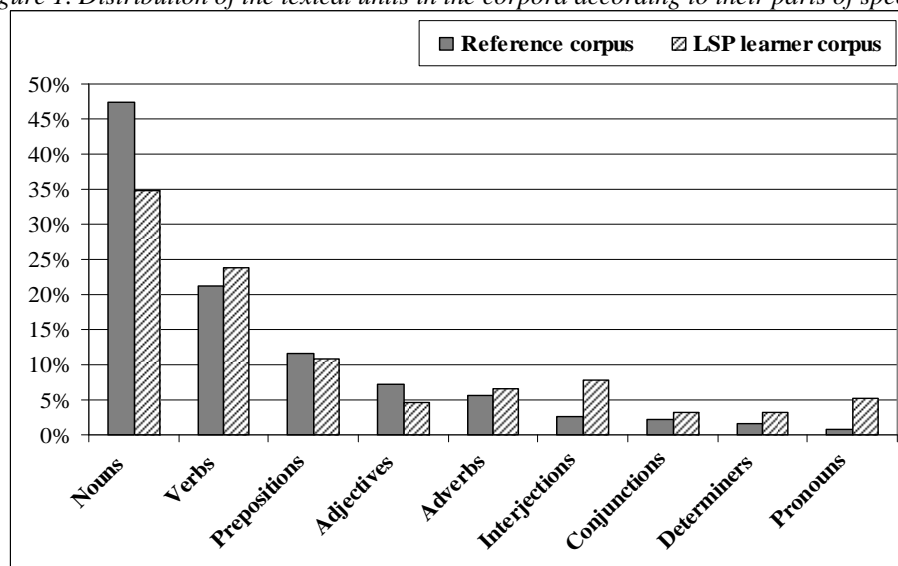
Trying to uncover and identify some of the specific features of learners' English radiotelephony requires a very detailed comparative analysis of the two corpora. As a starting point, an observation of the distribution and use of the modals occurring in the corpora is presented here.

Modals are an instructive example of the disparities which prevail between the corpora under investigation. Although they constitute about 4% of the verb tokens of each corpus, the modals occurring in the LSP learner corpus are much more diverse than the set of three modals found in the reference corpus: *will* and *should* account respectively for 82.5% and 16% of all the modal tokens of the reference corpus and *can* occurs only once (1.5%). The limited proportion of modals in the reference corpus illustrates one of the main features of phraseology and correlates with its specific characteristics of preciseness and concision. In the LSP learner corpus, however, up to seven different types of modals are used[7]: the most frequent, *can* and *will*, account respectively for about 37.5% and 34% of the modal tokens. The other 5 modals – *would*, *could*, *should*, *may* and *might* – respectively account for less than 10% of all the modal tokens in the LSP learner corpus. The distribution (in percentages) of these modals in the corpora is presented in the following table:

	will	would	can	could	may	might	shall	should	must
Reference corpus	82.5	0	1.5	0	0	0	0	16	0
LSP learner corpus	34	9	37.5	7	9	1.5	0	2	0

Table 2. Distribution (%) of the modal verbs in the corpora under study[8]

---

[7] *Must* and *shall*, being *hapax legomena* in the LSP learner corpus, have been excluded from our analysis.

[8] For the purpose of information only, we can specify that the overall distribution of the modals occurring in our spoken LSP learner corpus is rather different from the distribution of modals in spoken "natural" English, as determined by Kennedy's (2002) work on the distribution of modal verbs in the British National Corpus (BNC).

The analysis of the distribution of all the modals occurring in the LSP learner corpus and those occurring in the reference corpus will be carried out in the near future. For now, we will take a closer look at the distribution of the modal *can*, which only occurs, in the reference corpus, in a controller's interrogative message to convey a meaning of physical ability or capacity:

    1. C: […] can you accept departure from taxiway sierra 3? 1800 metres available.

The meaning of capacity conveyed by *can* in the reference corpus – as well as the exclusive semantic function of willingness conveyed by *will* and the one of weak obligation expressed by *should* – reflects the intention of the phraseology's designers to implement a direct and unambiguous language for specific purposes: "[…] expressions conveying contrasts in mood (degrees of certainty, possibility, necessity, futurity, etc.) as well as intentions and attitudes, are clearly incompatible with an aeronautical environment where precise, concise instructions, with a reference to present time are the rule" (Rubenbauer 2009: 51-52). In prescribed English phraseology, the imperative is the medium for conveying instructions and the auxiliary verb *can* is only used by controllers to verify the capacity of pilots to comply with forthcoming clearances: "Since the controller must ascertain, prior to issuing an instruction, that the pilot is actually in the position to carry out the instruction, the value of the modals is not one of politeness, but one of capacity ("CAN" = "(BE) ABLE TO")" (Rubenbauer 2009: 52).

Yet, in some contexts found in the LSP learner corpus, *can*, the most frequently used modal which occurs as much in declarative utterances as in interrogative ones, is often used otherwise. Its distribution seems to be related to the respective roles of the speakers involved in the communication: pilots mainly use *can* in interrogative sentences so as to make special requests regarding the optimisation of their flight, due to their role as users of the provided control services (examples 2 to 5); controllers, on the other hand, mainly employ *can* in declarative sentences to convey clearances, due to their role as administrators who provide pilots with manoeuvre instructions and authorisations (examples 2 to 5). In declarative contexts, *can* is used to a large extent by LSP French learners (controllers) to express permission or authorisation (79% of controllers' declarative utterances including the modal *can*) while the imperative form should be the privileged means to convey such notions.

    2. C: [...] welcome to you, can climb flight level 3 8 0 [...].
    3. C: [...] -huh- you can start for runway 2 7 Left, [...].
    4. C: okay […], continue approach OMAKO, -huh- you can fly heading 3 2 5.
    5. C: […] you can contact Marseille 1 3 3 decimal 8 8 0, bye bye.

In such contexts, the value of *can* is not one of physical or material capacity, but one of civility. Controllers seem to use *can*, rather than imperative structures, to courteously convey the expression of authority and obligation associated with the instructions and clearances they provide pilots with. This assumption is corroborated by the use of the modal *may* in similar contexts:

In the BNC spoken texts, *will* accounts for 26.5%, *would* for 21.5%, *can* for 23.1%, *could* for 9.4%, *may* for 2.3%, *might* for 3.9%, *shall* for 1.3%, *should* for 5.7% and *must* for 2.8%.

6. C: […], -huh- maintain level 1 4 0 on reaching and you may keep high speed.
7. C: […] continue on this heading, you may cross the localizer, […].
8. C: […] you may contact Marseille 1 3 2 2 5 5, have a nice flight, bye.
9. C: roger, -huh- you may climb level 3 4 0 sir, climb level 3 4 0.

This distinctive use of *can* and *may* in the LSP learner corpus is one of several examples of the various discrepancies between the corpora under study. The occurrence of the modals *would*, *could*, *may* and *might* in the LSP learner corpus is another example. This type of usage obviously belongs to a register different than prescribed phraseology: the use of modal verbs – but also subject pronouns (as in examples 3 to 9), determiners (example 7) and sometimes prepositions (example 3) – does not correspond to the linguistic characteristics of phraseology but rather to plain language. In example 9, the controller's awareness of using modality in a message that should have presented the imperative form – and which might thus contribute to potential ambiguity and misunderstanding – could be the reason for his reformulation or self-correction from "you may climb level 3 4 0" to "climb level 3 4 0".

6. Pedagogical Applications

The long-term pedagogical applications of our research project are varied. Our study is, however, still in progress and has not yet resulted in "off-the-shelf LSP learner-corpus-informed pedagogical materials" (Granger & Paquot forthcoming). Yet, several pedagogical applications are possible and practicable.

First, the primary data of the LSP advanced learner corpus could be a very useful pedagogical tool for institutions, such as the ENAC, which are in charge of future controllers and pilots' training. For instance, various activities based on the recordings, such as general comprehension exercises or blank filling exercises, could be developed by professionals in charge of English radiotelephony teaching. Second, the results from the contrastive interlanguage analysis carried out between the reference corpus and the LSP learner corpus together with the observation of regular usage patterns, whether correct or incorrect, could help compile up-to-date materials to meet future controllers' needs and interests. For instance, as far as modals are concerned, English radiotelephony teachers could present the potential ambiguity conveyed by the use of certain modals in certain structures while insisting on the needs for pilots and controllers to revert to plain language only when necessary.

In any case, the presentation and description of the language used by English radiotelephony advanced learners, whether complying with prescribed phraseology or corresponding to a more natural language, will help learners become well aware of the difficulties related to the actual use of this language for specific purposes.

7. Conclusion

Although the part-of-speech tagging of the corpora indicates a general pattern of similarities – both are mainly constituted of nouns, then verbs, then prepositions – a

closer observation of the distribution of these grammatical categories and their proportions, within each corpus, reveals pronounced discrepancies. The noun and adjective categories are, for instance, underrepresented in the LSP learner corpus while the verb, interjection, determiner and pronoun categories are overrepresented. It is also worth noting that the proportion of adverbs, prepositions and adjectives used in the LSP learner corpus does not seem to be related to the linguistic specificity of the domain as, according to Johansson & Hofland (1990 [1989]), these three categories are similarly represented in the word-class distribution of the tagged LOB Corpus[9] (which is definitely not the case of the noun, verb, determiner, pronoun and interjection categories).

Differences between the corpora under study have also been observed from a lexical, semantic and syntactic point of view, as illustrated by the use of the modals *can* and *may* in the LSP learner corpus. The preliminary results of our analysis of the latter suggest a tendency from English radiotelephony learners to use plain language in situations for which phraseology is provided, despite the potential misunderstandings that can result from using plain language. A more detailed contrastive interlanguage analysis of the two corpora will help describe the language used by advanced learners of English radiotelephony and thus try and identify their second language specific features in order to develop pedagogical materials which meet future controllers's (and pilots') needs.

References

Direction Générale de l'Aviation Civile (DGAC). (2007 [2000]). *Phraséologie. Manuel de Formation à la Phraséologie à l'Usage de la Circulation Aérienne Générale*. Ministère de l'équipement, des transports et du logement, Direction Générale de l'Aviation Civile, Direction de la Navigation Aérienne.

Gilquin, G., Granger, S. & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6(4), 319-335.

Gilquin, G., Papp, S. & Belén, D.M. (eds). (2008). *Linking Up Contrastive and Learner Corpus Research*. Amsterdam & Atlanta: Rodopi.

Granger, S. & Gilquin, G. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In J. Mukherjee & M. Hundt (eds) *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam & Philadelphia: John Benjamins, 55-78.

Granger, S., Hung, J. & Petch-Tyson, S. (eds). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (Language Learning and Language Teaching 6). Amsterdam & Philadelphia: John Benjamins.

---

[9] For more information about the Lancaster-Oslo-Bergen (LOB) corpus, see Johansson & Hofland (1989), among others.

Granger, S. & Paquot, M. (forthcoming). *Language for Specific Purposes learner corpora*. http://sites-test.uclouvain.be/cecl/archives/GRANGER_PAQUOT_Forthcoming_Language_for_Specific_Purposes_Learner_Corpora.pdf (last accessed on 29 October, 2011).

International Civil Aviation Organisation (ICAO). (2010 [2004]). *Manual on the Implementation of ICAO Language Proficiency Requirements*. Doc 9835-AN/453.

International Civil Aviation Organisation (ICAO). (2007 [1990]). *Manual of Radiotelephony*. Doc 9432-AN/925.

Johansson, S. & Hofland, K. (1990 [1989]). *Frequency Analysis of English Vocabulary and Grammar: Based on the LOB Corpus. Vol. 1, Tag Frequencies and Word Frequencies.* New York: Clarendon Press & Oxford University Press.

Kennedy, G. (2002). Variation in the distribution of modal verbs in the British National Corpus. In R. Reppen, S.M. Fitzmaurice & D. Biber (eds) *Using Corpora to Explore Linguistic Variation*. Amsterdam & Philadelphia: John Benjamins, 73-90.

Lopez, S., Condamines, A., Josselin-Leray, A., O'Donoghue, M. & Salmon, R. (2011). Linguistic analysis of English phraseology and plain language in air-ground communications. In *Proceedings from the 2011 Air Transport Research Society (ATRS) Conference*. Sydney, 29 June- 2 July 2011.

Mell, J. (1992). *Étude des Communications Verbales entre Pilote et Contrôleur en Situation Standard et Non-Standard*. Doctoral dissertation, University of Toulouse le Mirail.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (ed.) *Corpora and Language Teaching* (Studies in Corpus Linguistics 33). Amsterdam & Philadelphia: John Benjamins, 203-230.

Myles, F. (2008). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In L. Ortega & H. Byrnes (eds) *The Longitudinal Study of Advanced L2 Capacities*. Hillsdale: Lawrence Erlbaum, 58-72.

Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (ed.) *How to Use Corpora in Language Teaching* (Studies in Corpus Linguistics 12). Amsterdam & Philadelphia: John Benjamins, 125-152.

Philps, D. (1989). *L'Anglais de la Circulation Aérienne*. Toulouse: ENAC.

Philps, D. (1991). Linguistic security in the syntactic structures of air traffic control English. *English World-Wide* 12(1), 103-124.

Mis en forme : Français (France)

Rubenbauer, F. (2009). *Linguistics and Flight Safety: Aspects of Oral English Communication in Aviation*. Aachen: Shaker Verlag.

Tono, Y. (2000). A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In L. Burnard & T. McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective*. Papers from the Third International Conference on Teaching and Language Corpora, volume 2. Frankfurt am Main & New York: Peter Lang, 123-132.