Corpus linguistic methods

Evangelia Adamou

Adamou E. (2019). Corpus linguistic methods. J. Darquennes, J. Salmons & W. Vandenbussche (eds). ***Language contact***. Boston & Berlin: Mouton de Gruyter.

Corpus linguistic methods

Evangelia Adamou
French National Center for Scientific Research (CNRS)

Keywords: corpus linguistics; corpus-driven; contact linguistics; language contact; multilingual corpora; bilingual; annotation; codeswitching

1. Introduction

Corpus linguistics is a methodological approach that takes an empirical stance to the study of language. It relies on the analysis, whether qualitative or quantitative, of a body of written texts or transcriptions of spontaneous or semi-spontaneous speech. Corpus linguistic methods have a potentially strong impact on theory as they can offer support or challenge theoretical assumptions. They can also help improve our understanding of previously described linguistic phenomena and can reveal new topics of investigation that had hitherto gone unnoticed. Moreover, corpus linguistics is closely related to various fields of applied linguistics, for example through the elaboration of pedagogical tools.

More specifically, in contact linguistics, a corpus-driven approach based on ecologically valid data allows for the examination of the constraints and social significance of bilingual speech. In addition, in experimental approaches to bilingualism, natural corpus data are used as basic frequency data in combination with the controlled data which are produced in a laboratory environment.

2. Overview of existing corpora and their relevance to contact linguistics

The use of corpus linguistic methods in the various sub-fields of linguistics largely depends on practical issues related to the availability of corpora or the ease with which a corpus can be built and searched. This section presents a brief overview of existing corpora and the ways in which they can be explored for language contact.

2.1. Written corpora

Written corpora have become fairly easy to compile thanks to the widespread use of computers and the internet. As a result, large electronic corpora of formally written speech, often comprising literary and journalistic texts, are currently available online for a variety of languages. For example, several million-word corpora exist even for small national languages (such as Albanian[1]) and minority languages (such as Basque[2] and Soviet Romani[3]). Electronic corpora aim to provide users with the tools to conduct online or offline searches, not only for

[1] Accessed at http://web-corpora.net/AlbanianCorpus/search/
[2] Accessed at http://www.ehu.eus/en/web/eins/egungo-testuen-corpusa-etc-
[3] Accessed at http://web-corpora.net/RomaniCorpus/search/

word frequency but also for collocations, concordances, and n-grams following the model of well-developed English corpora.[4]

Despite this considerable progress in corpus compilation, sizeable corpora from a variety of speakers and sources are still restricted to a small number of major communication languages. Anand, Chung and Wagers (2011) observe that 85% of the data that are part of the Linguistic Data Consortium[5] represent only five languages, i.e.: English, Chinese, Arabic, Spanish, and Japanese.

From a contact linguistic perspective, language contact phenomena are relatively limited in formal written texts of national languages. In contrast, informal written texts, compiled from social media and text messages (SMS), may provide information on a wider range of contact phenomena, such as codeswitching.

2.2. Spoken corpora

Spoken corpora are of special interest to linguists focusing on contact as they potentially include a wide range of contact phenomena and may reveal information about the cognitive processes involved in the production and comprehension of bilingual speech. Large spoken corpora of adult language have been constituted on the basis of scripted or unscripted speech from TV and radio broadcasting (an example in point is the British National Corpus).[6] Some smaller corpora provide transcripts and recordings of free-speech conversations (e.g., Santa Barbara corpus of Spoken American English[7] and Switchboard Corpus).[8] Most of these corpora, however, do not target interactions between bilinguals and can at best provide evidence for the study of borrowings.

Linguists interested in bilingualism and its short- and long-term effects have compiled new corpora involving adult bilingual speakers, whether from (post-)colonial settings, traditionally bilingual communities, or migrant communities. In order to document the most ordinary way of speaking recordings of unscripted speech were made either in natural conditions or during interviews, preferably conducted by community members. These recordings combine with corpora from scripted speech and transcripts of TV and radio broadcasting. However, we note that in most of the cases these corpora are not made available for the members of the scientific community.

2.3. Spoken corpora from endangered languages

A promising contribution of corpus linguistics to the study of language contact phenomena may be its growing use in language documentation initiatives. Indeed, corpora from endangered languages are increasingly being compiled and made available online. Examples include the Endangered Language Archive (ELAR),[9] the DoBeS Archive (Documentation of Endangered Languages),[10] the PARADISEC Archive (Pacific and Regional Archive for

[4] See for an example the BYU corpora, grouping the most widely used online corpora of English. Accessed at https://corpus.byu.edu/
[5] Accessed at https://www.ldc.upenn.edu/
[6] Accessed at http://www.natcorp.ox.ac.uk/
[7] Accessed at http://www.linguistics.ucsb.edu/research/santa-barbara-corpus
[8] See https://catalog.ldc.upenn.edu/ldc97s62
[9] Accessed at https://www.soas.ac.uk/elar/
[10] Accessed at http://dobes.mpi.nl/

Digital Sources in Endangered Cultures),[11] the AILLA Collection (The Archive of the Indigenous Languages of Latin America)[12] and the Pangloss Collection.[13]

2.4. Sign language corpora

In the past decades, research on sign languages has been on the increase, typically relying on video corpora of signed interactions (see, e.g., the Sign Language corpus at the Language Archive).[14] As summarized in Adamou et al. (in press), there are three types of research studies on bilingual sign language: (i) research on sign bilingualism that focuses on the study of bilinguals who use a sign language and a spoken language in its written form, (ii) research on bimodal bilingualism of hearing individuals who use a sign language and a spoken language, and (iii) research on code-switching that looks at bilinguals who use more than one sign language.

2.5. Bilingual acquisition corpora

Spoken language corpora have held a privileged position for both first and second language acquisition studies. Bilingual acquisition has been explored through methods ranging from parental diaries to extensive recordings of natural speech. Examples can be found in the Child Language Data Exchange System (CHILDES).[15]

2.6. Learner corpora

Spoken or written learner corpora such as the International Corpus of Learner English (ICLE)[16] are increasingly used to explore the effects of an L1 on an L2 (or an L3). Learner corpus research not only contributes to theories of second language acquisition, but also plays an important role in applied linguistics through the elaboration of new pedagogical tools (see Granger, Gilquin and Meunier 2015).

2.7. Parallel corpora

The term "multilingual corpora" covers what is also known as "parallel corpora", consisting of a body of texts translated in different languages which are aligned across languages, e.g. the European Parliament Proceedings Parallel Corpus.[17] Parallel corpora are mainly used for research in contrastive linguistics and translation studies. They also have great impact in translation training and in the development of translation tools for language learners. The Multilingual Student Translation project, which started in 2016, aims to compile a new, large multilingual student translation corpus that will bring together the two research paradigms, i.e.: learner corpus research and corpus-based translation studies.[18]

[11] Accessed at http://www.paradisec.org.au/
[12] Accessed at http://www.ailla.utexas.org/site/welcome.html
[13] Accessed at http://lacito.vjf.cnrs.fr/pangloss/index_en.html
[14] The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen, https://corpus1.mpi.nl/ds/asv/;jsessionid=B59DACD6D374DE413552402F9EBD5440?0
[15] Accessed at http://childes.psy.cmu.edu/
[16] Accessed at https://www.uclouvain.be/en-cecl-icle.html
[17] Accessed at http://www.statmt.org/europarl/
[18] See https://uclouvain.be/en/research-institutes/ilc/cecl/must.html

3. Transcription tools and annotation

In contact linguistics, corpus-based research is a rapidly growing field, but the development of bilingual corpora is associated with specific challenges. Indeed, despite considerable progress in the development of automated techniques, bilingual corpora remain extremely challenging for Natural Language Processing technologies, i.e.: automatic speech recognition, parsing, machine translation, information retrieval and extraction, and computational processing (see Çetinoğlu, Schultz and Vu 2016; Guzmán et al. 2016). This means that manual coding is still needed in order to solve issues that arise from the competing grammars and phonology of the languages in use. As a result, building a large bilingual corpus is very time-consuming.

It therefore becomes urgent for linguists to embrace collaborative approaches to linguistic data compilation,[19] for example through the online annotation of open data, including through crowdsourcing. Such initiatives can be encouraged by taking into consideration the creation of a corpus in the career of a linguist, based on a set of quality standards and citations (see discussion in Thieberger et al. 2015). Of course, providing access to the raw or annotated data on which corpus-based research relies raises ethical questions. In order to make primary or annotated data accessible to the scientific community while ensuring the protection of personal data, one has to see to it that the privacy of the speakers is respected.[20] That can be done through anonymizing the parts of the sound files and transcripts that may include personal information.

From the perspective of open-access data, the annotation standards used for a corpus gain in significance. The Worldwide Web Consortium (W3C)[21] recommends the Extensible Markup Language (XML) and the Text Encoding Initiative (TEI) standards[22] for annotation standards. For example, in an XML file, content is delimited by tags and surrounded by angle brackets in order to describe a document and specify information about it, i.e.: <sentence>, in angle brackets, marks the beginning of a sentence and </sentence>, in angle brackets with a slash, marks its end.[23] Depending on the research questions, annotators may provide an orthographic transcription, parse the corpus for syntactic analysis, and annotate parts of speech (nouns, proper nouns, verbs, conjunctions, particles, etc.) through using Universal Parts Of Speech (POS) tags, for example.[24] For lesser-known languages a morpheme-by-morpheme transcription and translation (or *interlinearization*) is often provided together with the free translation at the sentence level. These levels (i.e.: words, morphemes, glosses) should also be delimited in the XML using the above-mentioned tags. Within these tags, one can add information about the language, e.g., xml:lang="fr" for French. For bilingual corpora, in particular, one could tag the words depending on the languages involved. One could, e.g., use tags such as L1 (the word is used in monolingual or bilingual contexts but can be

[19] Some web-based collaborative platforms and tools: Brat at http://brat.nlplab.org/; Webanno at https://webanno.github.io/webanno/; Atomic at http://corpus-tools.org/atomic/; Annis at http://corpus-tools.org/annis; IMS Corpus Workbench at http://cwb.sourceforge.net/.

[20] For Europe, see among others https://ec.europa.eu/digital-single-market/en/open-access-scientific-knowledge-0 and for language resources https://www.clarin-d.de/en/help/legal-information-platform

[21] The Worldwide Web Consortium, see https://www.w3.org/Consortium/

[22] The Text Encoding Initiative (TEI) is a consortium that develops and maintains annotation standards, see http://www.tei-c.org/index.xml. Another reference for standards is TalkBank, see https://talkbank.org/ Annotations may also be unified around a common format called Universal Dependencies http://universaldependencies.org

[23] See tutorials at http://www.tei-c.org/Support/Learn/tutorials.xml

[24] Accessed at http://universaldependencies.org/u/pos/

recognized clearly in this context as part of the L1), L2 (the word is used in monolingual or bilingual contexts but can be recognized clearly in this context as part of the L2), multiple or ambiguous (words in this context could be either L1 or L2), mixed (words are partially in both languages and they do not exist as such in either the L1 or the L2), foreign words (for words from an L3 or past-contact languages), unknown (the annotator(s) cannot determine the language and meaning of the word), or other (non-words, or in written corpora symbols, emoticons, etc.). For a speech corpus, one can also provide a phonetic and phonological transcription as well as a prosodic annotation, for example by segmenting the corpus into intonation units. The use of UNICODE (e.g. UTF-8) is strongly recommended[25] as it guarantees the recognition of all characters on the Internet and on any computer. In XML one can also tag the beginning and the end of an associated sound and video file (*time alignment*) using the angle brackets and adding the relevant content information, e.g., <AUDIO start="0.1106" end="3.7594"/>. Concerning sound files, the WAV format is preferred for long-term archiving whereas the MP3 format is a good option for online access.

Moreover, a vital step in the compilation of a corpus is the production and maintenance of the metadata. Metadata may include sociolinguistic information on the speakers, the researchers, the content of the document, and the date and place of recording, allowing for extra-linguistic factors to be analysed. Metadata may follow the Dublin Core specifications[26] as defined by the Open Language Archives Community (OLAC) or use fuller standards, such as the ISLE Metadata Initiative (IMDI) developed by the Max Plank Institute and the Component MetaData Infrastructure (CMDI)[27] initiated by CLARIN and linked to the CLARIN Concept Registry.[28] Arbil[29] is a practical metadata editor for both IMDI and CMDI standards (Withers 2012).

Researchers should opt for annotation tools that conform to these standards and that are compatible with one another. A variety of general manual annotation tools can be adapted to the study of language contact phenomena. A very popular tool is ELAN[30] which synchronizes the audio and video files with annotation tiers (Sloetjes and Wittenburg 2008). Language contact phenomena can be tagged in these tiers, depending on the researchers' needs, and searches can be conducted with ELAN's concordance tool. Besides ELAN, Anvil[31] is often used in sign language annotation (Kipp 2014). EXMARaLDA[32] is another widespread annotation tool that has been tested for multilingual research (Schmidt and Wörner 2014). CLAN and CHAT are tool sets native to language files in the CHILDES database. PHON is its phonological software programme that allows comparisons between target and produced phonological forms.

For phonetics, EasyAlign[33] is used to semi-automatically align phonetic annotations to transcriptions in PRAAT.[34] The WebMAUS[35] service (Kisler et al. 2012) is another tool that has successfully been tested for forced alignment including corpora from lesser-known

[25] http://www.unicode.org/
[26] http://dublincore.org
[27] http://www.clarin.eu/content/component-metadata
[28] The Data Category Repository (DCR), known as ISOCat, is currently transferred under https://openskos.meertens.knaw.nl/ccr/browser/
[29] http://tla.mpi.nl/tools/tla-tools/arbil/
[30] Created by the Max Planck Institute for Psycholinguistics at Nijmegen, Netherlands. Accessed at https://tla.mpi.nl/tools/tla-tools/elan/
[31] http://www.anvil-software.org/
[32] http://exmaralda.org/en/
[33] http://latlcui.unige.ch/phonetique/easyalign.php
[34] http://www.fon.hum.uva.nl/praat/
[35] At the Bavarian Archive for Speech Signals, http://clarin.phonetik.uni-muenchen.de/BASWebServices/

languages. A similar tool is SailAlign[36] that can accommodate long and noisy data as well as transcription errors.

More information on corpus linguistic methods and best practices is available in Lüdeling and Kytö (2008, 2009), O'Keeffe and McCarthy (2010), Durand, Gut and Kristofferson (2014), Biber and Reppen (2015), Kirk and Andersen (2016), Gries and Berez (2017).

4. Corpus-driven studies in contact linguistics

4.1. Brief overview

Corpora are currently exploited in contact linguistics both in a top-down and bottom-up fashion. Top-down approaches rely on theoretical models that were primarily elaborated with monolingual data and aim to test the validity of these models for bilingual data. That is the case for the Matrix Language Frame (MLF) model (Myers-Scotton and Jake 2017), which builds on the Speaking model in psycholinguistics (Levelt 1989), and the Minimalist approach to codeswitching (MacSwan 2016), which considers that the Minimalist Program (Chomsky 1995) should also successfully account for bilingual data. Corpus studies can test the generalizations and predictions made by these models. For example, Herring et al. (2010) and Parafita-Couto and Gullberg (2017) investigate the switches occurring between determiners and their nouns in Spanish-English, Welsh-English, and Papiamento-Dutch bilingual speech in order to evaluate competing theories of codeswitching, namely the MLF model and generative approaches. On the other hand, bottom-up approaches consider that higher levels of abstraction are shaped by language practices. Linguists in this line of research typically observe patterns in bilingual corpora and aim to explain these patterns by paying close attention to factors such as frequency and priming, as with variationist approaches (Poplack 1980; Poplack and Dion 2012; Torres Cacoullos and Travis 2016) and usage-based approaches (Backus 1992, 2015; Quick et al. 2017).

Corpora can be used both in a corpus-illustrated approach (i.e.: to provide occurrences of specific phenomena) or in a quantitative approach (i.e.: to provide frequency information and comparisons between speakers and texts). For quantitative analyses, appropriate statistical models need to be used depending, among other things, on the size of the corpora (see Nock et al. 2009; Newman, Baayen and Rice 2011; Tagliamonte and Baayen 2012; Gries 2015).

At present, corpus-driven studies in contact linguistics rely on data from three main types of contact settings: post-colonial and colonial settings, immigrant communities, and traditional contact settings involving two or more native languages. Table 1 presents a brief overview of such studies and shows that most of them focus on language pairs that include a European language such as Spanish, English, and French. Studies on language contact among immigrants and immigrant communities constitute another dynamic area of research in corpus linguistics. In contrast, contact involving traditional languages seems to be the least explored area in corpus linguistic studies.

[36] http://sail.usc.edu/old/software/SailAlign/

Table 1. Brief overview of corpus-driven studies in contact linguistics

Type of contact	Language pairs	Locations	References
Post-colonial and colonial settings	English-Spanish	U.S.A. (New York, Los Angeles, New Mexico)	Poplack 1980; Silva-Corvalán 1986; Torres Cacoullos and Travis 2016
	Quechua-Spanish	Bolivia, Ecuador	van Hout and Muysken 1994; Gomez Rendon 2008; Bakker and Hekking 2012
	Guaraní-Spanish	Paraguay	Gomez Rendon 2008
	Otomí-Spanish	Mexico	Gomez Rendon 2008
	Swahili-English	Kenya	Myers-Scotton 1993
	Hindi-English	India	Si 2010
	Chinese-English	mainland China and Hong-Kong	Wang and Liu 2013
	Welsh-English	Wales	Deuchar 2006
	French-English	Canada	Poplack and Dion 2012
	Arabic-French	Morocco	Nait M'Barek and Sankoff 1988
	Nkep (Oceanic)-Bislama (English lexifier creole)	Vanuatu	Meyerhoff 2014
	Aboriginal English/Kriol-Gurindji (Pama-Nyungan)	Australia	McConvell and Meakins 2005
Immigrants and immigrant communities	Turkish-Dutch	Netherlands	Backus 1992
	Romani-Finnish and Romani-Turkish-Greek	Greece, Finland	Adamou and Granqvist 2015
Traditional languages	Bora (Huitotoan)-Resígaro (Arawakan)	Colombian-Peruvian Amazon region	Seifart 2015

4.2. Case studies

Once a corpus has been collected and annotated, researchers may conduct a number of searches depending on their research questions and on the specificities of the corpus. Adamou (2016) and Guzmán et al. (2016) suggest that comparable codeswitching data are needed in order to understand how frequent codeswitching really is in natural human communication. Based on 17 bi- or multilingual corpora,[37] Adamou (2016) observes that most corpora contain

[37] Taking into account annotated Afro-Asiatic corpora accessed at http://corpafroas.huma-num.fr/Archives/ListeFichiersELAN.php Counts from Gomez Rendon (2008) for Guaraní (Paraguay), Quichua (Ecuador), and Otomí (Mexico) unpublished corpora. Slavic from Adamou et al. (2016) accessed at http://lacito.vjf.cnrs.fr/pangloss. Romani from Adamou and Granqvist (2015) partly available at

less than 5% word tokens from the L2, and more rarely, up to 20%-35% L2 word tokens (see Figure 1). These frequencies may also shed light to the language switching costs that have been noted in the experimental literature. Recently, Adamou and Shen (2019) and Johns, Valdés Kroff, and Dussias (2018) showed that language switching costs align with the codeswitching habits of the community as documented in corpus studies.

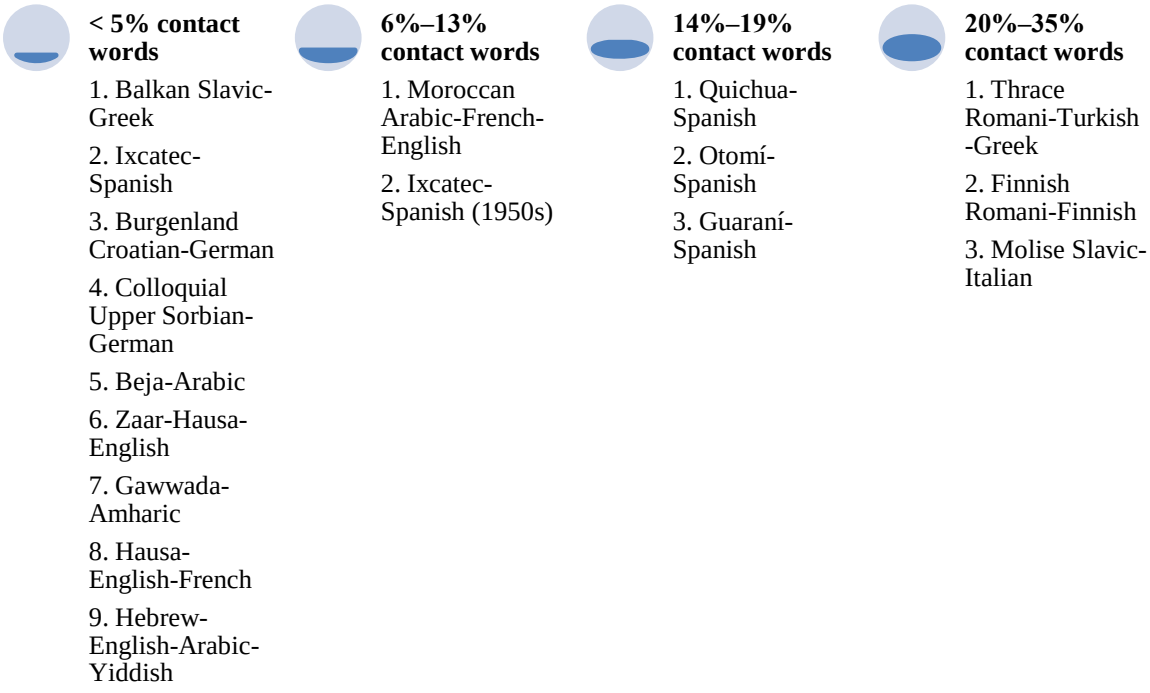

Figure 1. A scale of language mixing based on bilingual corpora (adapted from Adamou 2016)

The typological scale of language mixing presented in Figure 1 needs to be enriched by other layers of annotation. One topic of investigation is the degree of convergence between two languages in contact and its relation to codeswitching. Torres Cacoullos and Travis (2016), for example, present new corpus data from Spanish-English speakers from the traditionally bilingual community of New Mexico in the U.S. They show that, despite frequent codeswitching, pro-drop in the bilingual data is comparable to pro-drop in the monolingual varieties of Spanish and English. The reverse seems to be true for the Colloquial Upper Sorbian (Slavic)-German bilinguals who do not engage in frequent codeswitching with German and yet their pro-drop Slavic language has converged with non-prod-drop German (Adamou et al. 2016).

Pausing behaviour in relation to codeswitching is another research question that can only be addressed by making use of spoken corpora. Gardner-Chloros, McEntee-Atalianis, and Paraskeva (2013) recently showed that pauses do not correlate with codeswitching. They came to this conclusion on the basis of a study of two bilingual Greek Cypriot-English communities with different codeswitching habits: Cypriots in Great Britain for whom

http://lacito.vjf.cnrs.fr/pangloss. Ixcatec corpus, raw data at https://elar.soas.ac.uk/ and parts of the annotated corpus available at http://lacito.vjf.cnrs.fr/pangloss/corpus/list_rsc.php?lg=Ixcatec

codeswitching is the default mode of communication and Greek Cypriot speakers in Cyprus for whom codeswitching is a conversationally marked mode.

The existence of community patterns in codeswitching is another interesting topic that can only be tested with naturalistic data. Gardner-Chloros (2009), for example, notes that codeswitching is characterized by high variability across individuals. Adamou (2016), however, suggests that small bilingual communities may develop conventionalized ways of codeswitching. Codeswitching patterns were also noted in a study based on Bollywood scripts (Si 2010). This study shows that alternational codeswitching between Hindi and English has considerably increased in the past decades as opposed to previous codeswitching patterns that relied mainly on single-word insertions.

Finally, corpus studies allow us to move beyond anecdotal evidence for so-called "typologically-rare phenomena". Seifart (2015), for example, draws attention to previously unreported cases of affix borrowing in two Amazonian languages without any borrowing of the loanwords containing those affixes. Similarly, Romani corpora from Greece and Finland illustrate the regular insertion of L2 verbs with L2 verb morphology into L1-dominant speech (Adamou and Granqvist 2015) whereas in most codeswitching corpora contact verbs are regularly integrated into the morphology of the dominant language.

5. Conclusion

Corpus-driven research is a very dynamic paradigm in contact linguistics, as in other sub-fields of linguistics. This dynamism is due to the technological progress that enabled the growth of computerized corpora as well as the possibilities of online access to corpora, tools, and annotation platforms. Corpus linguistics is also particularly successful as it adequately addresses public and political concerns for increased accountability and replicability in science and fits the requirements of open science policies. In addition, corpus linguistics provides a solid empirical basis to test and shape theoretical models.

However, progress still needs to be made. The compilation of bi- and multilingual corpora remains a particularly time-consuming task as it requires considerable manual coding. Collaboration with Natural Language Processing specialists is therefore crucial for the development of new automated techniques adapted to the specificities of bi- and multilingual corpora. In the long term, increasing institutional support for FAIR data (i.e.: Findable, Accessible, Interoperable, and Re-usable data) and the development of web-based crowdsourcing will undoubtedly help to boost the production of bi- and multilingual corpora. The analysis of larger bi- and multilingual corpora from a wider variety of contact settings could then help improve our understanding of this important aspect of human communication.

References

Adamou, Evangelia. 2016. *A corpus-driven approach to language contact. Endangered languages in a comparative perspective*. Boston & Berlin: Mouton de Gruyter.

Adamou, Evangelia & Kimmo Granqvist. 2015. Unevenly mixed Romani languages. *International Journal of Bilingualism* 19(5). 525–547. doi.org/10.1177/1367006914524645

Adamou, Evangelia, Walter Breu, Lenka Scholze & Xingjia Rachel Shen. 2016. Borrowing and contact intensity: A corpus-driven approach from four Slavic minority languages. *Journal of Language Contact* 9(13). 515-544.

Adamou E. (2019). Corpus linguistic methods. J. Darquennes, J. Salmons & W. Vandenbussche (eds). ***Language contact***. Boston & Berlin: Mouton de Gruyter.

Adamou, Evangelia & Xingjia Rachel Shen. 2019. There are no language switching costs when codeswitching is frequent. *International Journal of Bilingualism* 23(1). 53–70. doi.org/10.1177/1367006917709094

Adamou, Evangelia, Onno Crasborn, Jenny Webster & Ulrike Zeshan. In press, 2020. Forces shaping sign multilingualism. In Ulrike Zeshan & Jenny Webster (eds.), *Sign multilingualism*. Berlin: De Gruyter Mouton & Nijmegen: Ishara Press.

Anand, Pranav, Sandra Chung & Matthew Wagers. 2011. Widening the Net: Challenges for Gathering Linguistic Data in the Digital Age. Submitted to the National Science Foundation SBE 2020 planning activity. https://www.nsf.gov/sbe/sbe_2020/2020_pdfs/Wagers_Matthew_121.pdf (accessed 16 January 2017)

Backus, Ad. 1992. *Patterns of language-mixing. A study in Turkish-Dutch bilingualism*. Wiesbaden: Harrassowitz.

Backus, Ad. 2015. A usage-based approach to codeswitching: The need for reconciling structure and function. In Gerald Stell & Kofi Yakpo (eds.), *Code-switching Between Structural and Sociolinguistic Perspectives,* 19–37. Berlin: Mouton de Gruyter.

Bakker, Dik & Ewald Hekking. 2012. Constraints on morphological borrowing: Evidence from Latin America. In Lars Johanson & Martine Robbeets (eds.), *Copies vs. cognates in bound morphology*, 187–220. Leiden: Brill.

Biber, Douglas & Randi Reppen (eds.). 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Çetinoğlu, Özlem, Sarah Schultz & Thang Vu. 2016. Challenges of computational processing of code-switching. In Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg & Thamar Solorio (eds.) *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, Austin, Texas, 1–11. Association for Computational Linguistics.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Deuchar, Margaret. 2006. Welsh-English code-switching and the Matrix Language frame model. *Lingua* 116(11). 1986–2011.

Durand, Jacques, Ulrike Gut & Gjert Kristofferson (eds.). 2014. *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press.

Gardner-Chloros, Penelope. 2009. *Code-switching*. Cambridge: Cambridge University Press.

Gardner-Chloros, Penelope, Lisa McEntee-Atalianis & Marilena Paraskeva. 2013. Codeswitching and pausing: an interdisciplinary study. *International Journal of Multilingualism* 10(1). 1–26.

Gomez Rendon, Jorge A. 2008. *Typological and social constraints on language contact: Amerindian languages in contact with Spanish*. Utrecht: LOT.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (eds.). 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2015. Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1). 93—117.

Gries, Stefan Th. & Andrea Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 379–409. Berlin & New York: Springer.

Guzmán, Gualberto A., Jacqueline Serigos, Barbara E. Bullock & Almeida J. Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg & Thamar Solorio (eds.), *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, Austin, Texas, 12–20. Association for Computational Linguistics.

Adamou E. (2019). Corpus linguistic methods. J. Darquennes, J. Salmons & W. Vandenbussche (eds). ***Language contact***. Boston & Berlin: Mouton de Gruyter.

Herring, John R., Margaret Deuchar, M. Carmen Parafita Couto & Monica Moro Quintanilla. 2010. 'I saw the madre': evaluating predictions about codeswitched determiner-noun sequences using Spanish-English and Welsh-English data. *International Journal of Bilingual Education and Bilingualism* 13. 553-573.

Johns, Michael A., Jorge R. Valdés Kroff & Paola E. Dussias. 2018. Mixing things up: How blocking and mixing affect the processing of codemixed sentences. *International Journal of Bilingualism.* First published on February 5, 2018. doi/10.1177/1367006917752570

Kipp, Michael. 2014. ANVIL: A Universal Video Research Tool. In Jacques Durand, Ulrike Gut & Gjert Kristofferson (eds.), *The Oxford Handbook of Corpus Phonology*, 420-436. Oxford: Oxford University Press.

Kirk, John M. & Gisle Andersen (eds.). 2016. Compilation, transcription, markup and annotation of spoken corpora. Special issue of the *International Journal of Corpus Linguistics* 21(3).

Kisler, Thomas, Florian Schiel & Han Sloetjes. 2012. Signal processing via web services: The use case WebMAUS. *Proceedings of Digital Humanities 2012*, 30–34.

Levelt, Willem J. M. 1989. *Speaking, from intention to articulation*. Cambridge, MA: MIT Press.

Lüdeling, Anke & Merja Kytö (eds.). 2008. *Corpus linguistics: An international handbook*, vol. 1. Berlin: Mouton de Gruyter.

Lüdeling, Anke & Merja Kytö (eds.). 2009. *Corpus linguistics: An international handbook*, vol. 2. Berlin: Mouton de Gruyter.

MacSwan, Jeff. 2016. Codeswitching in adulthood. In Elena Nicoladis & Simona Montanari (eds.), *Lifespan Perspectives on Bilingualism*, 183–200. Berlin: Mouton de Gruyter and Washington, DC: American Psychological Association.

McConvell, Patrick & Felicity Meakins. 2005. Gurindji Kriol: A mixed language emerges from code-switching. *Australian Journal of Linguistics* 25(1). 9–30.

Meyerhoff, Miriam. 2014. Borrowing in Apparent Time: With some comments on attitudes and universals. *Selected Papers from NWAV 42. University of Pennsylvania Working Papers in Linguistics* 20(2). 121–128.

Myers-Scotton, Carol. 1993. *Duelling Languages: Grammatical Structure in Code-switching*. Oxford: Clarendon press.

Myers-Scotton, Carol & Janice Jake. 2017. Revisiting the 4-M model: Codeswitching and morpheme election at the abstract level. *International Journal of Bilingualism* 21(3). 340–366/

Naït M'Barek, Mohammed & David Sankoff. 1988. Le discours mixte arabe-français: des emprunts ou des alternances de langue? *Revue Canadienne de Linguistique* 33. 143–154.

Newman, John, Harald R. Baayen & Sally Rice (eds.). 2011. *Corpus-based studies in language use, language learning, and language documentation.* Amsterdam: Rodopi.

Nock, Richard, Pascal Vaillant, Claudia Henry & Frank Nielsen. 2009. Soft memberships for spectral clustering, with application to permeable language distinction. *Pattern Recognition* 42. 43–53.

O'Keeffe, Anne & Michael McCarthy (eds.). 2010. *The Routledge Handbook of Corpus Linguistics.* London: Routledge.

Parafita-Couto, M. Carmen & Marianne Gullberg. 2017. Code-switching within the noun phrase: Evidence from three corpora. *International Journal of Bilingualism*, First published on September 14, 2017, https://doi.org/10.1177/1367006917729543

Poplack, Shana. 1980. Sometimes I'll start a sentence in Spanish y termino en espanol. *Linguistics* 18(7). 581-618.

Adamou E. (2019). Corpus linguistic methods. J. Darquennes, J. Salmons & W. Vandenbussche (eds). ***Language contact***. Boston & Berlin: Mouton de Gruyter.

Poplack, Shana & Nathalie Dion. 2012. Myths and facts about loanword development. *Language Variation and Change* 24(3). 279-315.

Quick, Endesfelder Antje, Elena Lieven, Malinda Carpenter & Michael Tomasello. 2017. Identifying partially schematic units in the code-mixing of an English and German speaking child. *Linguistic Approaches to Bilingualism*, First published on March 7, 2017, http://dx.doi.org/10.1075/lab.15049.qui

Schmidt, Thomas & Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut & Gjert Kristofferson (eds.), *Handbook of Corpus Phonology*, 402-419. Oxford: Oxford University Press.

Seifart, Frank. 2015. Direct and indirect affix borrowing. *Language* 91(3). 511-531.

Sloetjes, Han & Peter Wittenburg. 2008. Annotation by category - ELAN and ISO DCR. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Marrakech.

Si, Aung. 2010. A diachronic investigation of Hindi-English code-switching, using Bollywood film scripts. *International Journal of Bilingualism* 15(4). 388-407.

Silva-Corvalán, Carmen. 1986. Bilingualism and language change: The extension of *estar* in Los Angeles Spanish. *Language* 62(3). 587-608.

Tagliamonte, Sali & Harald R. Baayen. 2012. Models, forests, and trees of York English: *Was-were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1-21.

Torres Cacoullos, Rena & Catherine Travis. 2016. Two languages, one effect: structural priming in spontaneous code-switching. *Bilingualism: Language and Cognition* 19(4). 733-753.

van Hout, Roeland & Pieter Muysken. 1994. Modelling lexical borrowability. *Language Variation and Change* 6(1). 39–62.

Wang, Lin & Haitao Liu. 2013. Syntactic variations in Chinese–English code-switching. *Lingua* 123(1). 58–73.

Withers, Peter. 2012. Metadata Management with Arbil. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 72-75. Istanbul.