RiCL Research in Corpus Linguistics

Koder – A multi-register corpus for investigating register variation in contemporary German

Andressa Costa
PUC São Paulo / Brazil

Abstract – This paper introduces the design decisions in building the Koder corpus, a multi-register-corpus of contemporary German. The purpose of this corpus is to serve as a basis for the investigation into the use of German across registers. In order to construct a representative corpus, the essential considerations are: the type and number of registers to include, the number of texts in each register and minimal text length. The paper describes which aspects were central in determining these issues as well the corpus composition and the necessary text processing.

Keywords – corpus design; Koder; register; German

1. INTRODUCTION

The availability of corpora facilitates the investigation of language use considerably. At present, there are various German corpora available to the academic community. In spite of this, building a corpus is sometimes still necessary because they are not completely suitable for answering some research questions. This paper describes the design decisions and composition of Koder (*Korpus deutscher Register*). The purpose of this corpus is to serve as a basis for empirical investigations of the German language through different registers. Most studies look at linguistic phenomena only in one register or they investigate only spoken, written documents or documents from the Internet. Therefore, available corpora from German are neither diversified in terms of mode nor cover a wide range of registers. Nevertheless, materials from available corpora from the Institute for the German Language (IDS), *Dortmunder-Chat-Korpus* (Beißwenger 2013) and *German Political Speeches* (Barbaresi 2012) were integrated in this corpus.

Research in Corpus Linguistics 7: 69–83 (2019). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.07.04

The necessity for building this corpus comes from the intention to investigate some linguistic phenomena across different registers because, as Biber and Conrad (2009: 6–7) observe, the use of linguistic features is influenced by the register in which they are being used. Register, as used in this study, refers to "a variety associated with a particular situation of use (including particular communicative purpose)" (Biber and Conrad 2009: 6).

A central aspect to consider when building a corpus is representativeness. It involves determining the corpus size, that is, the number and types of texts to be included in the corpus, the number of words per text and the total number of words in the corpus as well as the types of registers, in the case of a multi-register corpus (Berber Sardinha 2004: 24–25). The decisions made on these aspects will depend on the goals of the analysis. However, more important considerations than corpus size are a definition of the target population and choices concerning the method of sampling (cf. Biber 1993a). In fact, Biber (1993a: 243) defines representativeness as "the extent to which a sample includes the full range of variability in a population."

As Biber (1993b: 219) notes, two kinds of error must be minimised to achieve a representative corpus: 'random error' and 'bias error'. A random error occurs when a sample is not large enough to accurately estimate the right population; a bias error is when the selection of a sample is systematically different from the population. Thus, one important consideration relates to how to sample language in a corpus to study general language. On this issue, Biber (1993b: 220) argues that "analyses must be based on a diversified corpus representing a wide range of registers to be appropriately generalised to the language as a whole." He justifies this view with the assumption that there is no adequate overall linguistic characterisation of an entire language; instead, there are marked linguistic differences across registers. In order to select the registers that adequately represent a language, it is necessary to consider the users of that language. Regarding corpus size, as Sinclair (2005) states, there is no maximum size. The author considers two main factors in establishing the minimum size of a corpus: "1. the kind of query that is anticipated from users; and 2. the methodology they [the researchers] used to study the data." To analyse linguistic variation using a corpus-based approach, Biber (1990) provides an empirical investigation with the following methodological issues regarding corpus construction:

1) How long texts should be to reliably represent the distribution of linguistic features in particular text categories;
2) How many texts within each text category are required to reliably represent the linguistic characteristics of that category and related questions concerning the validity of register categories;
3) How many texts are needed in a corpus to accurately identify the salient parameters of variation among texts;
4) How much of a cross-section is required to identify and analyse the salient parameters of variation among texts.

In his investigation, Biber (1990: 261–268) analyses and compares samples of different sizes using statistical techniques. The results indicate the following:

– There is a high level of stability for the analysed linguistic features in 1,000-word sub-samples of texts so that 2,000-word and 5,000-word texts in the standard corpora are reliable representatives of their text categories;
– 10-text sub-samples accurately represent the linguistic characteristics of register categories, including both the central tendency and the range of variation;
– A factor-analysis with a corpus of 120 texts and another with a corpus of 240 texts containing the full range of registers included in the original corpus (23 registers) reasonably well represents the underlying parameters of variation that were found in the initial factor analysis with a corpus of 481 texts;
– A corpus of 169 texts, with fewer registers than the other two samples, provides a poorer representation of the underlying parameters found in the original corpus.

Biber (1990: 269) showed in this study that "the underlying parameters of text-based linguistic variation […] can be replicated in a relatively small corpus if that corpus represents the full range of variation." Berber Sardinha (2004) applied the methodological procedures suggested by Biber (1990, 1993a) and proposed the minimum number of approximately 5,500,000 words for a general corpus of English and nearly 91,000 for a specific corpus. For their investigation of register variation in Brazilian Portuguese, Berber Sardinha *et al*. (2014) built a multi-register-corpus of 48 registers with 20 texts per register and texts with at least 400 words. The decisions made in designing Koder were based on the works of Biber (1990, 1993a, 1993b) and Berber Sardinha (2004) for determining the number of registers, as well the number of texts

within the register and minimal text length. The register selection was based on the typology developed in several chapters in Brinker *et al*. (2000) and Eroms (2008).

2. KODER (KORPUS DEUTSCHER REGISTER)

2.1. Register selection

The first step for register selection was to identify which registers are productive and represent the range of situational variation in contemporary German. This was not an easy task because there is no source where this information can be found. However, the typology of fields of communication presented in several chapters in Brinker *et al.* (2000) and Eroms (2008) for written and spoken texts served as a starting point. Brinker *et al.* (2000: XXVI) define fields of communication as an 'ensemble' of text types that are situationally and socially defined. This definition is to some extent similar to the definition of register adopted by Biber and Conrad (2009: 5), who consider register a category of texts with shared situational characteristics, whereas dialects are defined as a category of texts with shared social characteristics. The term 'field of communication' is not yet established, as Adamzik shows (2016: 126). Nevertheless, it was useful information to begin the selection of the register for this project.

The list of fields of communication proposed in Brinker *et al*. (2000), though comprehensive, has been considered provisional and unsystematic because an adequate typology for German texts has yet to be established. Moreover, it comprises only fields for written communication, excluding computer-mediated communication. Documents from the Internet and other registers like movies and non-fiction, which are not on the proposed typology, were added to this project. The selection of internet registers was based on Beißwenger and Lemnitzer (2013) and Berber Sardinha (2014), and the selection of movies on Veirano Pinto (2013). The second step was to determine the amount of registers. Because this corpus is currently being used as a basis for investigations about the general use of German and about individual linguistic features, the decision was made to include the complete range of registers described by the consulted literature.

Certain registers were selected from sources other than the consulted literature. This is the case for the registers under the label 'others' and the label 'oral communication' which comprises two categories from the *Database for Spoken*

German (DGD): *Forschungs- und Lehrkorpus* (FOLK) and *Gesprochene Wissenschaftssprache* (GWISS). Other registers from this database included in the corpus are conversations, oral exams and academic lectures. Material from Facebook and Twitter was collected from public accounts rather than from private users. The transcripts from TED talks subtitles were edited manually because they are automatically generated and contain many errors.

The registers included in this corpus represent a broad range of communicative situations in contemporary German, to which German speakers are currently exposed. It is not only diversified in terms of registers but also in terms of mode: the collection comprises written and spoken texts, as well as texts produced in a digital environment.

2.2. Text collection and corpus size

After the selection of registers, text size and the number of texts had to be determined. For this purpose, two aspects were taken into consideration (Biber 1990: 258):

1. How many texts within each text category are required in order to represent the linguistic characteristics of that category reliably and the validity of register categories;
2. How long texts should be in order to reliably represent the distribution of linguistic features in a particular text category.

The first decision made was to build a balanced corpus in which all registers have the same number of texts. In order to determine how many texts each register should contain other studies using multi-register corpora served as orientation (Biber 1988; Biber *et al.* 2006; Xiao 2009; Berber Sardinha *et al.* 2014). Most of these studies did not use a balanced corpus except for Berber Sardinha *et al.* (2014), who used a corpus composed of 20 texts per register and included texts with at least 400 running words. The decision made for Koder was to collect 50 texts of at least 400 words for each register in order to build a corpus as large as possible in a limited time.

Nevertheless, some registers have more than 50 texts, whereas others have fewer than that. The reason why some registers, such as TED talks, detective series, and academic lectures, have fewer than 50 texts lies in the difficulty to find enough available material. Other registers have fewer than the minimum number of words established as part of the corpus design criteria (at least 400 words per text). Because

several texts from news, recipes, readers' letters to the editor, job advertisements, and song lyrics have fewer than 400 words, more texts were added to reach the minimum word length. Except for job advertisements and news, the following criteria were settled for the addition more texts:

- Recipes: 50 dishes were selected and two or more different recipes for each dish were collected;
- Readers' letter to the editor: 50 editions from magazines and newspapers were selected and all readers' letters to the editor were collected;
- Song lyrics: 50 singers or bands were selected and three songs from each singer or band were collected.
- Some internet registers have the same problem regarding text length. The decision in this case was the following:
  - Twitter: sets of tweets from about 50 different hashtags;
  - Facebook comments: sets of comments from 50 different posts;
  - YouTube comments: sets of comments from 50 different videos;
  - Reader commentary: sets of comments from about 50 different articles;
  - Wikipedia user talk: sets of comments from the editors of about 50 different Wikipedia articles.

The first purpose of this corpus is to serve as a basis in an investigation on register variation through the multi-dimensional approach (Biber 1988) in which a factor analysis is conducted in order to identify which linguistic features significantly co-occur in the specific registers. A pilot study undertaken to test the data revealed that the sample size (Kaiser-Meyer-Olkin measure of sampling adequacy = .84) is very good for conducting a factor analysis. Thus, Bartlett's Test of Sphericity (< .0001) shows that the correlation between the variables in the data is significantly different from 0, which means that they are suitable for a factor analysis (Loewen and Gonulal 2015: 187–188).

2.3. Text selection and compilation

The decision about which texts to compile depended upon the availability of the materials. This criterion includes both available corpora and the permission to use material found on the Internet. Firstly, a list of text types was made on the basis of the

literature.[1] Subsequently, a search for available corpora was made and the texts of these corpora were collected. Afterwards, the availability of other text types to be collected without any legal restrictions was checked. Most of the texts were collected from the Internet and some of them were scanned.

Documents were collected from available corpora as follows: conversation, institutional communications and interviews were collected from FOLK; oral exams and academic lectures were collected from GWISS; Wikipedia user talk were compiled from *Deutsches Referenzkorpus* (DeReKo); professional chats from the *Dortmunder-Chat-Korpus* (Beißwenger 2013) and *German Political Speeches* is a corpus developed by Barbaresi (2012). The FOLK and GWISS corpora as well DeReKo are provided by *Institute for the German Language* (IDS). Material of the majority of registers was completely compiled from the Internet except for material of editorial and readers' letters to the editor which were partially scanned and partially compiled from the Internet.

The compilation of texts from the Internet involved the following criteria: a) a survey was undertaken in order to list newspapers, magazines, publishers, institutions, companies, websites about recipes, blogs, etc. from Germany and with the domain *.de*; b) the author of documents, such as academic texts and articles from newspapers and magazines as well as fictional literature, had to be German. When the author's origin could not be checked, the text was discarded. However, it was difficult to apply these criteria to Tweets and commentaries from Facebook and YouTube. In this case, the material was still collected. It is important to note here that the data from these three registers was gathered from public profiles. No data from personal profiles was collected.

The register academic and scientific institutions comprises two sub-registers: academic texts and popular science. Academic texts contain three different text types but only doctorate theses are split into groups: one group is composed of documents from Human and Social Sciences, the other group of documents from Natural, Engineering and Biological Sciences. In the collection, there are exclusively academic articles from Human and Social Sciences because it is difficult to obtain academic articles from Natural, Engineering and Biological Sciences written in German. It seems to be a tendency in such disciplines to write articles in English rather than in German. In

[1] See Tables 1 and 2 below.

contrast, popular science from Natural, Engineering and Biological Sciences articles which are written in German could be easily found. Academic textbooks are extracts which could only be found by one publisher. There is not much material available on the Internet: 38 texts are from Human and Social Sciences and 12 texts from Natural, Engineering and Biological Sciences. Similar to academic textbooks, the texts from fictional literature and non-fiction are extracts compiled from the websites of different German publishers.

Documents from media registers were selected from different national newspapers and magazines except for spoken news and news. Spoken news was collected from a German broadcaster which provides the transcriptions of the news on the website. The category news, which comprises short news, was collected from regional newspapers from different regions in Germany.

For the compilation of song lyrics, the following criteria were adopted: a) 50 singers and bands were selected from hit lists; b) research about the artists was made in order to select three songs by each artist which were composed between 1990 and 2018. The music genres are diverse: pop, rock, hip hop and rap.

The selection of movies and series occurred in two phases. Firstly, a list of German movies and series was made through a search on the web; secondly, a search for subtitles of the listed movies and series was conducted. The final selection contains the material which could be found. The variety of German series and films could not be successfully represented in this corpus because of a lack of available subtitles.

After the selection of texts described in the forerunning, the corpus content is summarised in Table 1 and Table 2. The documents are grouped into two categories: written (Table 1) and spoken registers (Table 2). The registers are categorised according to their fields of communication. Some registers have various text types, which are also identified in the tables. Moreover, the texts are classified according to features as dialogue/monologue, scripted/non-scripted, public/private, etc.

Mode	Fields of Communication	Register	Sub-registers	Setting	Specific text-type included in corpus	Texts	Number of words
WRITTEN	University and Scientific fields (Heinemann 2000a)	Academic texts	Academic articles (Human and Social Sciences)	Public	Article	50	287,052
			Academic textbook (Human and Social Sciences)	Public	Textbook	32	164,517
			Academic textbook (Natural Sciences and Engineering)	Public		7	29,144
			Academic textbook (Biological Sciences)	Public		11	55,523
			Doctoral thesis (Human and Social Sciences)	Public	Doctoral thesis	50	3,298,682
			Doctoral thesis (Natural Sciences and Engineering)	Public		39	1,339,736
			Doctoral thesis (Biological Sciences)	Public		11	332,142
		Popular science	Specialised/Technical texts (Natural Sciences and Engineering)	Public	Article	27	51,010
			Specialised/Technical texts (Biological Sciences)	Public		23	47,862
	Medicine and Health (Wiese 2000)	Package insert		Public	Manual	50	91,608
	Political Institutions (Klein 2000)	Plenary minutes		Public	Minute	50	551,938

Table 1: Written section of Koder (*Korpus deutscher Register*)

WRITTEN	Documents from the web (Beißwenger and Lemnitzer 2013; Berber Sardinha 2014)	Blog	Professional chat	Public		50	45,087
		Website		Public		50	36,003
		Wikipedia article		Public		50	253,394
		Wikipedia user talk		Interactive /Public		50	77,905
		Chat		Interactive /Public		50	100,395
		Facebook comments		Interactive /Public		50	39,749
		YouTube comments		Interactive /Public		50	123,664
		Tweets		Interactive/Public		50	452,165
		Reader commentary		Interactive /Public		50	155,823
	Mass Media (Burger 2000)	Reader's letter to the editor		Public	Letter	70	57,843
		News		Public	News	100	31,704
		Newspaper article		Public	Article	50	47,097
		Editorial		Public	Editorial	50	38,504
		Magazine article		Public	Article	50	57,843
		Commentary/opinion		Public		50	46,346
	Economics and Commerce (Hundt 2000)	Business communication		Public	Invitations/ Letters/	50	132,182
	Jurisprudence and the Legal System (Busse 2000)	Legal texts from the school and university		Public	Schools laws/ Resolutions/ Regulations	50	279,433
	Everyday Use (Heinemann 2000b)	Recipe		Public	Manual	200	40,957
		Instruction manual		Public		50	184,978
		Horoscope		Public		50	40,626
		Job advertisements		Public		105	30,271
	Fiction Literature (Eroms 2008)	Prose		Public	Novel/romance	50	85,341
				Public	Youth literature	50	274,865
	Other	Non-fiction		Public		50	207,243
Total Written						**1,875**	**9,088,632**

Table 1 (continuation)

Mode	Fields of Communication	Register	Sub-registers	Setting	Specific text-type included in corpus	Texts	Number of words
SPOKEN	Everyday Use (Heinemann 2000b)	Conversation	Pop/Rock/Pop-Rock/Hip Hop/Rap	Dialogue/Private		50	499,322
		Song lyrics		Monologue/Scripted		150	49,369
	Other	Institutional communication (FOLK)		Dialogue/Public		50	319,111
		Interview (FOLK)		Dialogue		23	205,662
		TED talk		Monologue/Non scripted		30	53,055
	Church and Religion (Simmler 2000)	Sermons		Monologue/ Scripted		50	169,535
	Mass Media (Burger 2000)	Spoken news		Monologue/ Scripted		50	30,560
		Newspaper interview		Monologue/ Scripted	Interview	50	87,701
	University and Scientific fields (Heinemann 2000b)	Academic speech	Oral exam	Monologue/Non-scripted	Exam	38	167,735
			Experts' Lecture (manuscript)	Monologue/Scripted		50	356,987
			Experts' Lecture (transcription)	Monologue	Lecture	14	87,032
			Students' Lecture (transcription)	Monologue		29	134,862
	Political Institutions (Klein 2000)	Political speech		Monologue Scripted		50	90,271
	Fiction Literature (Eroms 2008)	Theatre		Dialogue scripted		50	279,286
		Films (Veirano Pinto 2013)	Documentary	Monologue/Dialogue scripted		50	253,428
			Series	Dialogue scripted	Drama, Comedy, Romance	38	110,808
			Films	Dialogue scripted	Detective	23	78,130
				Dialogue scripted	Comedy	50	367,376
				Dialogue scripted	Drama	50	281,587
Total Spoken						**897**	**3,621,817**

Table 2: Spoken section of Koder (*Korpus deutscher Register*)

An important remark to be made is that this corpus is intended to be a monitor corpus. The design decisions described here refer to this first version which will be used to investigate register variation in contemporary German. More texts and registers will be added in the future according to necessity.

2.4. Processing

After the material was collected, some processing was needed. All the text files were converted into text (.txt) format, either manually, by copying and pasting, or automatically, with the command pdftotext for PDF files. Sometimes, PDF files had to be manually corrected when they were converted into text format because the content became unreadable. Some transcripts from subtitles had an *l* instead of an *I* and vice versa in words as *Ich* which was then written *lch* or *als* which was written *aIs*. Contracted forms with *'s* were normalised to *es*. All these corrections were made using *sed*, a Unix utility which can be used for editing data (Kochan and Wood 2016: 70).

The files were cleaned semi-automatically: most texts had to be cleaned manually because the material to be removed was not uniform across texts. The texts from the spoken registers, chat and Wikipedia user talk had uniform material so that it was possible to use scripts written in Shell to clean them. The scripts are listed in Appendix 1.

3. CONCLUSION

This paper presented Koder, a multi-register-corpus of contemporary German which is composed of diversified register categories from a broad range of communicative situations. Thus, it comprises written and spoken texts as well as texts from computer mediated communication. The building of this corpus required decisions to be made not only about the number of texts and words but also about the number and type of registers to be included. The selection of the register categories was based on fields of communications (Brinker *et al.* 2000; Eroms 2008) and expanded with the addition of more registers from specific domains such as the Internet (Beißwenger and Lemnitzer 2013; Berber Sardinha 2014), films (Veirano Pinto 2013) and available corpora (DeReKo, *Folk and Gwiss, Dortmunder-Chat-Korpus* and *German Political Speeches*).

Decisions regarding the number of texts and words were guided by works of Biber (1990, 1993a, 1993b) and Berber Sardinha (2004).

The Koder corpus comprises a broad range of register categories which are used in the most diverse communicative situations by German speakers. Notwithstanding, some essential categories such as different types of television programmes could not be included in this first version of the corpus due to time and technical limitations. The expansion of the corpus in terms of registers, sub-registers and number of texts as well as a conducting further research will be considered in future research.

REFERENCES

Adamzik, Kirsten. 2016. *Textlinguistik. Grundlagen, Kontroversen, Perspektiven*. Berlin: Mouton de Gruyter.

Barbaresi, Adrien. 2012. *German Political Speeches, Corpus and Visualization*. http://purl.org/corpus/german-speeches (15 November, 2016.)

Beißwenger, Michael. 2013. *Das Dortmunder Chat-Korpus: Ein Annotiertes Korpus zur Sprachverwendung und Sprachliche Variation in der Deutschsprachigen Chat-Kommunikation. LINSE*. http://www.linse.uni-due.de/tl_files/ PDFs/ Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf (22 December, 2018.)

Beißwenger, Michael and Lothar Lemnitzer. 2013. Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt "Digitales Wörterbuch der deutschen Sprache" (DWDS). *Journal for Language Technology and Computational Linguistics* 26/2: 1–22.

Berber Sardinha, Tony. 2004. *Lingüística de Corpus*. Barueri: Manole.

Berber Sardinha, Tony. 2014. 25 years later: Comparing Internet and pre-Internet registers. In Tony Berber Sardinha and Márcia Veirano Pinto eds., 81–105.

Berber Sardinha, Tony, Carlos Kaufmann and Cristina Acunzo. 2014. Dimensions of register variation in Brazilian Portuguese. In Tony Berber Sardinha and Márcia Veirano Pinto eds., 35–79.

Berber Sardinha, Tony and Márcia Veirano Pinto eds. 2014. *Multi-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5/4: 257–269.

Biber, Douglas. 1993a. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 242–257.

Biber, Douglas. 1993b. Using register diversified corpora for general language studies. *Computational Linguistics* 19/2: 219–241.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, Douglas, Mark Davies, James K. Jones and Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1/1: 1–37.

Brinker, Klaus, Gerd Antos, Wolfgang Heinemann and Sven Sager eds. 2000. Preface. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., XXIII–XXVIII.

Brinker, Klaus, Gerd Antos, Wolfgang Heinemann and Sven Sager eds. 2000. *Linguistics of Text and Conversation: An International Handbook of Contemporary Research*. *Volume 1*. Berlin: Mouton de Gruyter.

Burger, Harald. 2000. Textsorten in den Massenmedien. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 614–628.

Busse, Dietrich. 2000. Textsorten des Bereichs Rechtwesen und Justiz. In Klaus Brinker Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 658–675.

Deutsches Referenzkorpus (DeReKo), Wikipedia Diskussionen 2015. http://corpora.ids-mannheim.de/pub/wikipedia-deutsch/2015/ (20 November, 2017)

Eroms, Hans-Werner. 2008. *Stil und Stilistik: Eine Einführung*. Berlin: Schmidt.

Heinemann, Margot. 2000a. Textsorten des Bereichs Hochschule und Wissenschaft. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 702–709.

Heinemann, Margot. 2000b. Textsorten des Alltags. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 604–614.

Hundt, Markus. 2000. Textsorten des Bereichs Wirtschaft und Handel. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 642–658.

IDS, *Datenbank für Gesprochenes Deutsch* (DGD), FOLK. http://dgd.ids-mannheim.de (9 October, 2019.)

IDS, *Datenbank für Gesprochenes Deutsch* (DGD), GWSS. http://dgd.ids-mannheim.de (9 October, 2019.)

Klein, Joseph. 2000. Textsorten in Bereich politischer Institutionen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 732–755.

Kochan, Stephen and Patrick Wood. 2016. *Shell Programming in Unix, Linux and OS X* (fourth edition). Indiana: Addison-Wesley.

Loewen, Shawn and Talip Gonulal. 2015. Exploratory factor analysis and principal components analysis. In Luke Plonsky ed. *Advancing Quantitative Methods in Second Language Research*. London: Routledge, 182–212.

Simmler, Franz. 2000. Textsorten des religiösen und kirchlichen Bereichs. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 676–690.

Sinclair, John. 2005. Corpus and text – Basic principles. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1–16.

Veirano Pinto, Márcia. 2013. *A Linguagem dos Filmes Norte-americanos ao Longo dos Anos: Uma Abordagem Multidimensional*. São Paulo: PUC São Paulo dissertation.

Wiese, Ingrid. 2000. Textsorten des Bereichs Medizin und Gesundheit. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann and Sven Sager eds., 710–718.

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28/4: 421–450.

Corresponding author
Andressa Costa
Praca Marechal Deodoro, 60
01150-010 São Paulo, Brazil
e-mail: acosta.andressa@gmail.com

received: August 2019
accepted: October 2019

Appendix 1: Scripts used for the automatic cleaning of some texts

EVERY DAY CONVERSATION and INSTITUTIONAL COMMUNICATION

```
cat filename.txt | grep [A-Za-z] | cut -f3 | sed -e 's/(.)//g' -e 's/([0-9]\.[1-9]*//g' -e 's/°hhh//g' -e 's/°hh//g' -e 's/flüstert//g' -e 's/°h//g' -e 's/°hh//g' -e 's/°hhh//g' -e 's/hhh°//g' -e 's/hh°//g' -e 's/h°//g' -e 's/räuspert sich//g' -e 's/räuspert//g' -e 's/lacht//g' -e 's/lachen//g' -e 's/+++//g' -e 's/schmatzt//g' -e 's/schnalzt//g' -e 's/seufzt//g' -e 's/hustet//g' -e 's/schluckt//g' -e 's/unverständlich//g' -e 's/schnieft//g' -e 's/Lachansatz//g' -e 's/Gemurmel,//g' -e 's/Gemurmel//g' -e 's/Blättern,//g' -e 's/Gelächter,//g' -e 's/Gelächter//g' -e 's/Geräusche,//g' -e 's/kichert//g' -e 's/stöhnt//g' -e 's/Nebengeräusche,//g' -e 's/Nebengeräusche//g' | tr -d '[]()' |

grep -v '^$'> filename_clean.txt
```

INTERVIEW, ORAL EXAM and LECTURE

```
cat filename.txt | grep [A-Za-z] | cut -f3 | sed 's/(.)//g' | sed 's/([0-9]\.[1-9]*//g' | sed 's/°hhh//g'| sed 's/°hh//g' | sed 's/°h//g' | sed 's/hhh°//g' |sed 's/hh°//g' | sed 's/h°//g' | sed 's/räuspert sich//g' | sed 's/räuspert//g' | sed 's/lacht//g' | sed 's/+++//g' | sed 's/schmatzt//g' | sed 's/schnalzt//g' | sed 's/hustet//g' | sed 's/schluckt//g' | sed 's/unverständlich//g' | sed 's/lachend//g' | tr -d '[]()' | grep -v '^$'> filename_clean.txt
```

CHAT

```
cat chattxt/ filename | grep -A1 '<messageBody>' chattxt/pc45.txt | tr '<' '\n' | grep -v '>' | cut -d'-' -f4 | grep -v '^$' >filename_clean.txt
```

WIKIPEDIA USER TALK

```
cat wd_txt/filename | grep -A1 '<p>' wd_txt/file | tr '<' '\n' | tr -d '[]()' | grep -v '>' | cut -d'-' -f4 | grep -v '^$' >wd_clean/filename_clean.txt
```