TGermaCorp – A (Digital) Humanities Resource for (Computational) Linguistics

Andy Lücking, Armin Hoenen, Alexander Mehler
Goethe University Frankfurt
Text Technology Lab
luecking@em.uni-frankfurt.de, hoenen@em.uni-frankfurt.de, mehler@em.uni-frankfurt.de

Abstract

TGermaCorp is a German text corpus whose primary sources are collected from German literature texts which date from the sixteenth century to the present. The corpus is intended to represent its target language (German) in syntactic, lexical, stylistic and chronological diversity. For this purpose, it is hand-annotated on several linguistic layers, including POS, lemma, named entities, multiword expressions, clauses, sentences and paragraphs. In order to introduce TGermaCorp in comparison to more homogeneous corpora of contemporary everyday language, quantitative assessments of syntactic and lexical diversity are provided. In this respect, TGermaCorp contributes to establishing characterising features for resource descriptions, which is needed for keeping track of a meaningful comparison of the ever-growing number of natural language resources. The assessments confirm the special role of proper names, whose propagation in text may influence lexical and syntactic diversity measures in rather trivial ways. TGermaCorp will be made available via `hucompute.org`.

Keywords: German literature resource, language diversity, corpus characteristics, linguistic annotation

1. Introduction

TGermaCorp is a digital humanities resource build around German literature texts from several centuries.[1] The primary texts are annotated on four levels: Firstly, the parts of speech are tagged according to the STTS (Schiller et al., 1999). Secondly, each token is assigned to its lemma. Thirdly, proper names are classified according to the kind of their referent (e.g., person or institution). Fourthly, clauses, sentences, paragraphs and headings are explicitly marked. Fifthly, multiword expressions are identified. All annotations have been carried out by linguistically trained annotators.

One characteristic of TGermaCorp is the composition of its primary sources: TGermaCorp is designed in view of capturing the lexical and morpho-syntactic varieties of written German as exhibited in German-speaking literature. Thus, TGermaCorp complements corpora that address the homogeneous style of mainly contemporary German (e.g., newspaper texts) like *TIGER* (TIGER project, 2003), (Brants et al., 2004), *DeReKo* (DEREKO project, 1999), (Dipper et al., 2002), or the *Huge German Corpus* (Stuttgart, 2010), (Schiller et al., 1999), as well as individual language resources like the *Kant Korpus* (Schmitz and Stark, 2008) (Lenders and Schmitz, 2007). Given these characteristics, TGermaCorp aims at applications and investigations within the field of *Digital Humanities* and therefore is located in the low-resource intersection area between computational linguistics and the study of literature. The corpus will be made publicly available via `www.hucompute.org` under the Creative Commons license CC BY-SA 3.0 DE.

2. Qualifying TGermaCorp

What are the primary sources of TGermaCorp and in which way have they been collected? What is the size of TG? Are the POS annotations of TGermaCorp reliable? These questions are addressed subsequently.

[1] 'TGermaCorp' is not an acronym though being related to "text technology", "German" and "corpus".

2.1. Composition of Primary Sources

In order to obtain a selection of literary texts, we followed canonical advices as documented in three expert sources: the LiMoST database for motives and themes[2], the canon of Marcel Reich-Ranicki,[3] and a guide for students of German philology (Segebrecht, 2006). We compiled the corpus from three freely available archives: Project Gutenberg[4] (PG), respectively the German partner Gutenberg-DE, WikiSource[5] (WS), and Deutsches Textarchiv[6] (DTA). All three sources provide texts that are free of copyright restrictions. Given their mentioning in at least one of the expert sources and their availability in one of the text archives, 238 texts from 107 authors have been selected. An excerpt of each text has been drawn. Excerpts are of various lengths (mean: 383 token, SD: 135 token). Due to their canonical salience, Johann Wolfgang von Goethe and Friedrich Schiller contributed most text excerpts (16 each). Texts stem from the 16th century until nearly today. Poems and theatrical plays are also included, as are German translations of non-German works. Poetic texts are designated as such and can easily be excluded from analyses since lyrical language use is special in several respects (Bierwisch, 2008). In addition to the excerpt compilation, TGermaCorp contains two complete texts, namely Thomas Mann's novel "Der Tod in Venedig" and the Wikipedia article on "Genetik"[7] (*genetics*). The latter is included as a sample of a text of different provenance. By this means, TGermaCorp aims at representing diversity of a single target language, which has to be dealt with in, e.g., philological studies and is often underestimated in natural language processing ap-

[2] `http://zs.gbv.de/motive/index.html`

[3] `https://de.wikipedia.org/wiki/Der_Kanon`, accessed multiply between May and December 2013

[4] `https://www.gutenberg.org/`

[5] `https://de.wikisource.org/`

[6] `www.deutschestextarchiv.de/`

[7] `http://de.wikipedia.org/wiki/Genetik`, accessed on 8th November 2012.

Table 1: Summary of concatenated POS.

POS	freq
ADV/ART	1
APPR/ART	25
APPR/PPER	1
ART/PPER	1
KOKOM/ART	1
KOUS/PPER	6
NN/ADJA	2
PIS/PPER	3
PPER/PPER	29
PRF/PPER	5
PWAV/ART	2
PWAV/PPER	2
PWAV/PRF	2
VAFIN/ART	1
VAFIN/PPER	38
VMFIN/PPER	13
VVFIN/PPER	30
VVIMP/PPER	1

Table 2: Frequencies of POS.

Rank	POS	freq	Rank	POS	freq
3	$,	10164	14	NE	2205
8	$.	7167	1	NN	20444
13	$(	2353	34	PAV	358
9	ADJA	6095	29	PDAT	527
11	ADJD	3890	33	PDS	369
5	ADV	7619	24	PIAT	949
43	APPO	92	25	PIS	942
7	APPR	7420	6	PPER	7499
19	APPRART	1346	15	PPOSAT	2137
47	APZR	55	51	PPOSS	23
2	ART	10641	44	PRELAT	72
32	CARD	385	22	PRELS	1027
38	FM	181	20	PRF	1237
36	ITJ	250	54	PROAV	1
28	KOKOM	706	45	PTKA	59
10	KON	4645	42	PTKANT	104
41	KOUI	105	26	PTKNEG	930
18	KOUS	1526	21	PTKVZ	1057
27	PTKZU	828			
49	PWAT	42			
30	PWAV	435			
31	PWS	413			
50	TRUNC	29			
12	VAFIN	3283			
53	VAIMP	14			
37	VAINF	232			
40	VAPP	122			
23	VMFIN	967			
48	VMINF	47			
52	VMPP	18			
4	VVFIN	7771			
35	VVIMP	347			
16	VVINF	2093			
39	VVIZU	167			
17	VVPP	1621			
46	XY	56			

plications. In order to provide a simple lexical example: the German noun *door*, *Tür*, is spelled "Thür" in Immermann's *Muenchhausen*; likewise the adverb *freilich* (*certainly*) is spelled "freylich" in von Schubert's *Ansichten von der Nachtseite der Naturwissenschaft*. Such spelling variations are only the tip of the iceberg – these and many more complications need to be addressed.

2.2. Some Facts and Figures

TGermaCorp comprises 122,902 word tokens. The average token length (excluding punctuations) is 4.59 characters, with a range of 1 to 39.[8] Tokens are assigned their parts of speech (POS) in terms of the *Stuttgart-Tübingen TagSet* (STTS) (Schiller et al., 1999). Note that we follow a "concatenation" approach to POS, based on the model of the APPRART tag. That is, word tokens that are contractions of two lexical units[9] are tagged with the concatenation of the POS involved. For instance, the token "kanns", which is a contraction of the modal verb "kann" (*can*) and the pronoun "es" (*it*), is tagged with VMFIN/PPER (we use a slash '/' as concatenation operator). In total there are 163 contractions in TGermaCorp, which are summarized in Table 1. However, since we do not assume that concatenated POS constitute proper parts of speech, we bifurcate them and use the split, "atomic" categories for analyses. The summary of the split POS is given in Table 2.

With regard to named entities, we basically followed the three classes *Person*, *Location* and *Organisation* used in the CoNLL 2003 training data set, which set a practical standard for named entity recognition.[10] However, since we didn't expect many organisations to be mentioned in literary texts and since the CoNLL threefold distinction ignores many further kinds of names, we distinguished between names of persons ("nepers"), names of locations ("neloc"), and other kinds of names ("nemisc") in the annotation of TGermaCorp. The latter are then assigned their specific kind, for instance, "chrononym" or "institutionym". The list of admissible kinds of names has been compiled from several resources (viz. Brendler (2004); Debus (2012); Kamianets (2000); Nagel (2008); Vasil'eva (2011), the *Urban Dictionary* (Urban Dictionary LLC, 1999), and *Wiktionary* (Wikimedia Foundation, Inc., 2009)), giving rise to a rather detailed inventory of proper name classifications. In sum, there are 1,586 names of persons, 347 names of locations, and 104 other kinds of names. The latter are mainly used to refer to mythological (*Mythonym*) or theological entities (*Theonym*), or to name rivers (*Potamonym*) or art objects (*Artionym*) – see Table 3 for the complete classification of other kinds of names.

[8] One-character tokens may result from interjections ("A!") or abbreviations – both are separated from their punctuation in the process of tokenisation.

[9] We found no token that consists of more than two items.

[10] See `http://www.cnts.ua.ac.be/conll2003/ner`; see also `http://nlp.stanford.edu/software/CRF-NER.shtml`.

Table 3: Classification of other kinds of names.

1	Theonym	34
2	Potamonym	14
3	Mythonym	11
4	Artionym	9
5	Institutionym	6
6	Pragmatonym	4
7	Synthroponym	4
8	Biblionym	3
9	Anthonym	2
10	Chrononym	2
11	Dokumentonym	2
12	Epochonym	2
13	Potejonym	2
14	Anchistonym	1
15	Hemeronym	1
16	Koilonym	1
17	Oikodomonym	1
18	Poetonym	1
19	Porejonym	1
20	Therionym	1
21	Titlonym	1
22	Urbanonym	1

Table 4: Overview: reliability assessment.

Tagset	PercAgree	Kappa	AC1	AC1 Conf.
STTS	77.12	0.87	0.87	(0.85 – 0.89)
UT	84.68	0.92	0.92	(0.91 – 0.94)

2.3. Reliability

In order to assess the reliability (Carmines and Zeller, 1979) of the part-of-speech annotation of TGermaCorp we calculated the interrater agreement of several annotators and different data. The main agreement study comprises five annotators' STTS annotations of an extract of 555 words of Thomas Mann's novel *Der Tod in Venedig*. Additionally, the STTS annotation has been mapped onto the 12 tags of the Universal Tagset (UT) (Petrov et al., 2012). Agreement has been measured by means of three coefficients: raw percentage agreement ("PercAgree"), Fleiss' Kappa (Fleiss, 1971), and Gwet's AC1 (Gwet, 2001). The respective results are collected in Table 4. The reliability results reach Krippendorff's (Krippendorff, 1980) level of credible results (i.e., agreement coefficient > 0.80), which, according to (Rietveld and van Hout, 1993) can even be regarded as "almost perfect". We used the R environment for statistical computing (R Core Team, 2013) for all analyses and calculations.

3. Assessing the Lexical and Syntactic Range of TGermaCorp

How does TGermaCorp diverge from related German corpora? To answer this question, we compute a number of diversity measures to compare TGermaCorp with two "reference corpora", that is, TigerSmall and WikiMimikry, sampled for this purpose.

3.1. TigerSmall and WikiMimikry

In order to obtain reference corpora of comparable size, we randomly sampled texts of equal size as texts in TGermaCorp starting from two third-party sources. The first comparison corpus is called *Wikipedia-based Mimikry Corpus* (WikiMimikry). It has been built by extracting the plain text of Wikipedia articles out of the German dump from 30th April 2015 using the *WikiExtractor*[11]. The second comparison corpus consists of sentences sampled from the *Tiger Treebank* (TIGER project, 2003).

For the purpose of comparing the corpora in a fair way, their plain texts have been POS tagged and lemmatized using one and the same preprocessing tool. We used the TreeTagger (Schmid, 1994) to this end, for which Giesbrecht and Evert (2009) report an overall accuracy of 95.82 on the TIGER treebank (given their specific application conditions). For syntactical analysis, we converted the TreeTagger output of the corpora to the CoNLL 2009 format[12] and parsed the result using the latest version of the MALT parser[13]. Thus, if there is some noise induced by the preprocessing procedure utilising the specific tools mentioned, all corpora should be affected in a similar way.

3.2. Lexical diversity

In order to assess the lexical diversity of the corpora, we computed their coverage with respect to the German release of *Wiktionary*[14] of 1st September 2015. This has been done on the level of wordforms and lemmas, excluding punctuations. On the level of wordforms, Wiktionary covers 87.00 % of TGermaCorp, 86.00 % of TigerSmall, and 82.00 % of WikiMimikry. On the level of lemmas, 60.49 % of TGermaCorp, 54.36 % of TigerSmall, and 48.50 % of WikiMimikry are covered.

Following Baayen (1992), quoted after Evert and Baroni (2007), we additionally calculated the "measure of productivity", viz. the portion of hapax legomena, as a further indicator of lexical diversity. The results are shown in Table 5.

Finally, we computed measures of type-token ratio (TTR) for both tokens vs. types (i.e., unique wordforms) and tokens vs. lemmas (classified for their POS). The results are summarized in Table 6.

Since the TTR is known to be dependent on contingent features like text length, we looked for other expressive measures for lexical richness (despite our three corpora being of approximately the same length). Following Covington and McFall (2010), we calculated MATTR (*moving average* TTR) as the average of TTR values observed in sliding windows of 500 tokens. This measure of lexical diversity does not depend on text length. The MATTR values are 0.72 for TigerSmall, 0.64 for the TGermaCorp and 0.62 for WikiMimikry.

[11] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[12] See http://ufal.mff.cuni.cz/conll2009-st/task-description.html

[13] http://www.maltparser.org/

[14] https://de.wiktionary.org

Table 5: Proportions of hapax legomena.

Source	prop. of hapax legomena	prop. of dislegomena
TGermaCorp	0.652	0.141
TigerSmall	0.657	0.136
WikiMimikry	0.641	0.137

Table 6: TTR: variants and results.

Source	WF token / WF type	WF token / lemma × POS
TGermaCorp	0.211	0.182
TigerSmall	0.239	0.204
WikiMimikry (TreeTagger pos-lemma)	0.214	0.188

3.3. Sentence Similarity and Syntactic Complexity

In order to assess corpus-internal similarity of sentences we perform a Monte-Carlo simulation on the comparison corpora. We start with randomly sampling 1000 sentences from each corpus. Then, we compute the Jaccard coefficient for each pair of these sentences, where each sentence is represented by the multiset of its wordforms. This procedure is iterated 1000 times for each corpus. Finally, the resulting similarity distributions are averaged and ranked per corpus. The results are shown in Figure 1(a).

Secondly, we adopt the method of measuring tree-like structures in social ontologies developed by Mehler (2011) for comparing parse trees of sentences generated by the MALT parser. We used, for example, the measure D_m of Altmann and Lehfeldt (1973), which recursively assesses the complexity of subtrees as a ratio of their widths and depths (Abramov and Mehler, 2011). We mapped each sentence of each sample of each corpus on a vector of 13 such measures of tree-like structuring (including D_m and 12 measures taken from Mehler (2011, chap. 3.4.1)). By a Monte-Carlo simulation (of 1000 iterations) we draw 1000 sentences from each corpus and computed for each sample the distances of these vectors. The resulting averaged rank distribution obtained by applying the Euclidean distance are shown in Figure 1(b). According to Euclidean distance, the Tiger Corpus seems to contain the least similar sentences. However, using a different distance measure, viz. the Mahalanobis distance, results are far more leveled – see Figure 1(c).

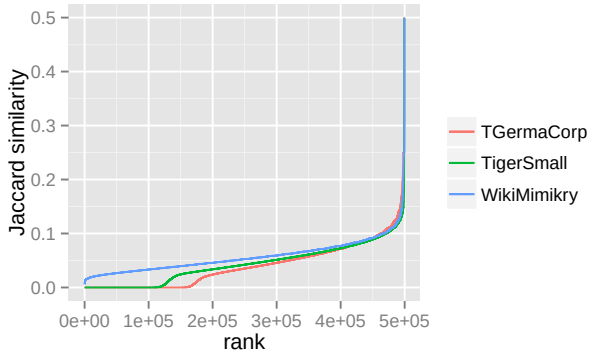
3.4. Choice of Measures

In general, comparability with external resources requires measures which are more general or widely used, such as the Euclidian distance. At the same time, different measures emphasize different aspects of the data. For instance, the Mahalanobis distance is the only distance of the three measures used that takes covariance into account. Whilst the other distances are thus insensitive to the context of the other vectors, the Mahalanobis distance compares each pair of sentences with respect to their position in the overall space of the sample, which in turn gives rise to a more levelled output. Using the Wilcox test (R library MASS (Venables and Ripley, 2002)), which consistently deals with ties for statistical significance on the distribution of all distances from the Monte Carlo simulations, we find that all pairwise distances are highly significant for the Jaccard similarities and the Euclidian distances. However, for the Mahalanobis distance between TGermaCorp and WikiMimikry, there was no significant difference. Since the distance measure determines the basis for further statistical and ultimately interpretative assessments, it is favorable to chose a number of different measures and/or to motivate the choice of measure carefully. The question for the effect of measure has been raised with respect to various subfields various times, see for instance Cha (2007) on density functions providing a dendrogram of distance measures, Salleh et al. (2012) on geometrical shapes, Cerqueira-Silva et al. (2009) on molecular markers. The latter reported a highly significant Spearman correlation of 0.58 between the Mahalanobis distance and the Euclidian distance, making them the most distant measures for their data and distance set. In computational linguistics, Rama and Kolachina (2012) worked on typological distances. Jin and Barrière (2005) found in a preliminary study, that the Dice coefficient, most similar to the Jaccard index, correlated best with human similarity judgments. Given these considerations, the choice of the three applied measures allows for the assessment of different aspects of the data and allows generalisability on the other hand.

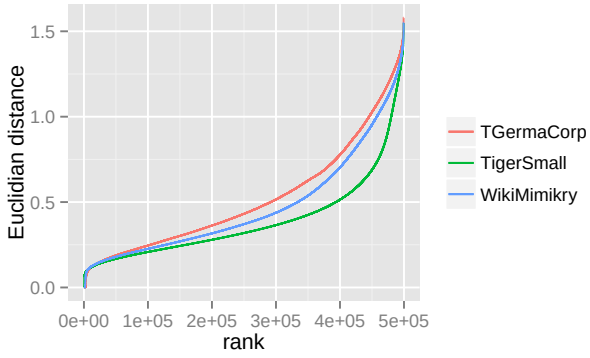
4. Discussion

Given the different genres underlying TGermaCorp and the two comparison corpora, the quite similar results of the diversity measures applied above come as a surprise. However, those measures focus on the respective span of the feature in question within a target corpus while ignoring mutual overlap. This line of reasoning is fostered by observing that the Wiki articles that make up the WikiMimikry comparison corpus contain a lot of named entities (which are furthermore written in a vast variety of typescripts, including Greek, Chinese, and so on) which are not part of the vocabulary of the other resources – see Table 7 for respective figures.

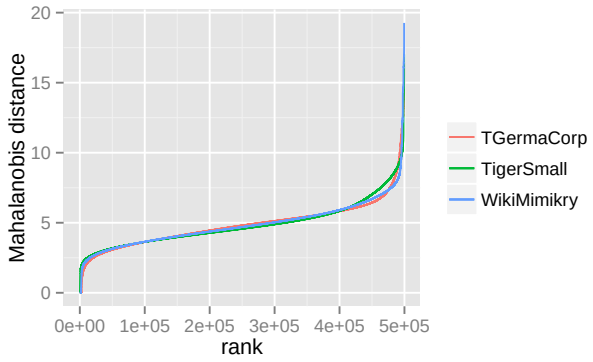
With this in mind, we also calculated a simple mutual lexical overlap between the comparison corpora on the level of lemmas as identified by the TreeTagger. As can be



(a) Multi-set similarity (Jaccard similarity).



(b) Tree complexity (Euclidean distance).



(c) Tree complexity (Mahalanobis distance).

Figure 1: Comparing corpora quantitatively (rank-distance plots of Monte-Carlo simulations).

Table 7: Frequency of NEs within the comparison corpora based on TreeTagger outcome.

Resource	Number of NEs
TGermaCorp	2410
TigerSmall	6159
WikiMimikry	9208

Table 8: Lexical overlap (lemma).

	TGermaCorp	TigerSmall	WikiMimikry
TGermaCorp	—	4586	4548
TigerSmall		—	6778
WikiMimikry			—

seen from the results reported in Table 8, WikiMimikry and TigerSmall are most alike in terms of a shared vocabulary, while TGermaCorp is approximately as distinct from TigerSmall as from WikiMimikry.

Note that the results are not adjusted for named entities (recall Table 7): a comparison solely in terms of proper content and function words leads to an decrease of the lexical overlap between TigerSmall and WikiMimikry which is proportionally larger than the respective decreases observed for TGermaCorp – cf. Table 9.

Table 9: Lexical overlap (lemma) excluding NEs.

	TGermaCorp	TigerSmall	WikiMimikry
TGermaCorp	—	4383	4306
TigerSmall		—	6159
WikiMimikry			—

5. Conclusion

This paper introduced TGermaCorp as a novel resource especially devoted to the computational analysis of literary data. We described the sampling and annotation of the texts of TGermaCorp and provided a quantitative comparison regarding two reference corpora – drawn from Wikipedia and from the Tiger treebank. TGermaCorp can be used to train NLP tools that are better adapted to literary data (not being addressed by Tiger). Our assessment shows that in terms of lexical similarity of sentences and their complexity the TGermaCorp and the Tiger treebank are comparable. However, part of the diversity is due to the influence of proper names, which occur with different frequencies in various resources. Accordingly, assessing lexical overlap provides quantitative evidence for the fact that TGermaCorp contains historical texts whose vocabulary is not in the focus of present-day language resources like Wiktionary. Furthermore, given the proliferation of natural language resources, quantitative assessments of the kind employed in order to

characterize TGermaCorp are useful for comparing corpora and eventually pinpoint their specific features.

6. Acknowledgements

We thank our student assistants Laura Becker, Lena Franz, Theresa Kasperkowitz, Elena Panina, and Sarah Richter (in alphabetical order) for their work in annotating the TGermaCorp.

7. Bibliographical References

Abramov, O. and Mehler, A. (2011). Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4):291–336.

Altmann, G. and Lehfeldt, W. (1973). *Allgemeine Sprachtypologie (Prinzipien und Messverfahren)*. Wilhelm Fink Verlag, München.

Baayen, H. (1992). Quantitative aspects of morphological productivity. In Geert Booij et al., editors, *Yearbook of Morphology 1991*, Yearbook of Morphology, pages 109–149. Springer Netherlands.

Bierwisch, M. (2008). Linguistik, Poetik, ästhetik. *Zeitschrift für Literaturwissenschaft und Linguistik*, 38(150):33–55.

Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Journal of Language and Computation*, 2:597–620.

Brendler, S. (2004). Klassifikation der Namen. In Andrea Brendler et al., editors, *Namenarten und ihre Erforschung. Ein Lehrbuch für das Studium der Onomastik*, chapter 2, pages 69–92. Baar, Hamburg.

Carmines, E. G. and Zeller, R. A. (1979). *Reliability and Validity Asessment*. Quantitative Applications in the Social Sciences. SAGE, Beverly Hills and London.

Cerqueira-Silva, C., Cardoso-Silva, C., ao, L. C., Nonato, J., Oliveira, A., and Corrêa, R. (2009). Comparison of coefficients and distance measurements in passion fruit plants based on molecular markers and physicochemical descriptors. *Genetics and Molecular Research*, 8:870–879.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Covington, M. A. and McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Debus, F. (2012). *Namenkunde und Namengeschichte. Eine Einführung*. Grundlagen der Germanistik. Erich Schmidt Verlag, Berlin.

Dipper, S., Kermes, H., König-Baumer, E., Lezius, W., Müller, F. H., and Ule, T. (2002). DEREKO – (DEutsches REferenzKOrpus) German reference corpus. Final report (part i). Technical report, IMS: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, SfS: Seminar für Sprachwissenschaft, Universität Tübingen.

Evert, S. and Baroni, M. (2007). *zipfR*: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, pages 29–32. (R package version 0.6-6 of 2012-04-03).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the German web as corpus. In *Proceedings of the fifth Web as Corpus workshop*, pages 27–35.

Gwet, K. (2001). *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, Gaithersburg, MD.

Jin, Z. and Barrière, C. (2005). Exploring sentence variations with bilingual corpora. *Corpus Linguistics Conference, Birmingham, United Kingdom*.

Kamianets, W. (2000). Zur Einteilung der deutschen Eigennamen. *Grazer Linguistische Studien*, 54:41–58.

Krippendorff, K. (1980). *Content Analysis*, volume 5 of *The SAGE KommTexT Series*. SAGE Publications, Beverly Hills and London.

Lenders, W. and Schmitz, H.-C. (2007). Die Elektronische Edition der Schriften Immanuel Kants. *Kant Studien*, 98(2):223–235.

Mehler, A. (2011). Social ontologies as generalized nearly acyclic directed graphs: A quantitative graph model of social ontologies by example of wikipedia. In Matthias Dehmer, et al., editors, *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, pages 259–319. Birkhäuser.

Nagel, S. (2008). *Lokale Grammatiken zur Beschreibung von lokativen Sätzen und ihre Anwendung im Information Retrieval*. Ph.D. thesis, Ludwig-Maximilians-Universität München.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC 2012, pages 2089–2096.

R Core Team, (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rama, T. and Kolachina, P. (2012). How good are typological distances for determining genealogical relationships among languages? In *Proceedings of the 24th International Conference on Computational Linguistics*.

Rietveld, T. and van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour*. Mouton de Gruyter, Amsterdam.

Salleh, S. S., Aznimah, N., Aziz, A., Mohamad, D., and Omar, M. (2012). Combining Mahalanobis and Jaccard distance to overcome similarity measurement constriction on geometrical shapes. *International Journal of Computer Science Issues*, 9(4).

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Segebrecht, W. (2006). *Was sollen Germanisten lesen?* Erich Schmidt Verlag, Berlin, 3 edition.

Vasil'eva, N. (2011). Die Terminologie der Onomastik, ihre Koordinierung und lexikographische Darstellung. In Karlheinz Hengst et al., editors, *Namenkundliche Informationen*, volume 99/100, pages 31–45. Leipziger Universitätsverlag, Leipzig.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

8. Language Resource References

DEREKO project. (1999). *DEREKO (Deutsches Referenzkorpus)*. IDS Mannheim, SfS Tübingen, IMS Stuttgart.

Hans-Christian Schmitz and Werner Stark. (2008). *Das Bonner Kant-Korpus*. Universität Bonn, distributed via Korpora.org.

IMS Stuttgart. (2010). *Huge German Corpus (HGC)*. IMS Stuttgart.

TIGER project. (2003). *TIGER Corpus*. Universität des Saarlands, Universität Stuttgart, Universität Potsdam, 2.2.

Urban Dictionary LLC. (1999). *The Urban Dictionary*. urbandictionary.com.

Wikimedia Foundation, Inc. (2009). *Wiktionary*. wiktionary.org.