The Challenges and Joys of Analysing Ongoing Language Change in Web-based Corpora: a Case Study

Anne Krause
Leipzig University
Beethovenstraße 15
04107 Leipzig
anne.krause@uni-leipzig.de

Abstract

Researchers of language variation and change often need to go to great lengths to find sufficient data, particularly when they shall be used for a sound statistical analysis of the phenomenon in question. The recent analogical change in the formation of the imperative singular of German strong verbs with vowel gradation is a case in point, as it could not have been studied without the compilation of a web-based corpus. On the one hand, the investigation was faced with a number of challenges during the compilation of the corpus, the search for relevant hits and their annotation for a number of variables. On the other hand, results which would otherwise not have been obtained balance out this increased amount of manual labour. The present paper elaborates on some of these challenges and provides suggestions how they might be avoided in similar investigations in future. It concludes by presenting invaluable insights which would not have been gained without the present corpus study.

1 Introduction

It has been noted several times by different authors that the use of the web as corpus enriches investigations of linguistic variation and change by providing a higher number of authentic and more recent examples than traditional corpora can furnish. In contrast to such “opportunistic” uses of the web, researchers of recent language change may be forced to make “systematic” use of web-based corpora because they are “the only source for examples of very rare usages and constructions” (Mair, 2012, 245). In the present project, a web-based corpus has been compiled as the primary source of evidence, not only because the web yielded more examples than traditional corpora but because the only text type which yielded enough evidence is specific to the web.

Instead of consulting a large pre-existing web-based corpus, material from a very specific website was used in the current investigation; nevertheless, problems faced during corpus compilation and analysis and suggestions for avoiding them can be generalised to similar phenomena and languages to a great extent.

1.1 Change-in-progress in German verb inflection

There are a number of German strong verbs which exhibit a stem vowel change from the infinitive -e- to -i- in the imperative singular, for example the verb *geben* ‘give’ in Table 1:

		present	
number	person	indicative	imperative
singular	1st	*geb(e)*	***gib***
	2nd	*gibst*	
	3rd	*gibt*	
plural	1st	*geben*	*gebt*
	2nd	*gebt*	
	3rd	*geben*	

Table 1: Conjugation table for the German verb *geben* ‘give’

The present project investigates the replacement of the established i-stem imperative singular of these strong verbs with vowel gradation by an e-stem variant formed in analogy to weak (regular) verb inflection, e.g. *sterben* ‘to die’: *stirb*! → *sterb(e)*!; *geben* ‘to give’: *gib*! → *geb(e)*!

Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, pages 27–34,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

1.2 The Conserving Effect

Along the lines of former usage-based analyses of analogical language change, it is hypothesised that the established i-stem imperative singular forms of lower frequency verbs are replaced by analogical variants earlier and faster than those of higher frequency verbs. For example, native speakers of German consistently stumble over the expression *Milk die Kuh!* 'Milk the cow!', employing the established i-stem imperative form of a low frequency verb, but they seem to accept both variants of verbs from a middle frequency region such as *bewirb dich!/ bewerb(e) dich!* 'apply (for sth.)!'. On the other end of the scale, the analogically formed variants of high frequency verbs in sentences like *Geb mir das Buch!* 'Give me the book!'/ *Seh es dir an!* 'Have a look at it!' are usually frowned at, whereas the i-stem variants of the same verbs are not.

The imperative singular forms of high frequency verbs are assumed to resist analogical change because they are highly entrenched in speakers' minds; this phenomenon is generally referred to as the "Conserving Effect" (Bybee and Thompson, 1997, 380). Although it has been explained from very early on that this frequency effect could be found in "modern leveling" as well (Hooper, 1976, 99), the majority of research in this area has been concerned with cases of completed language change. The present study thus fills two gaps by examining change-in-progress in German, a language in which the effects of type and token frequencies are still underresearched.

1.3 Imperative singular forms in corpora

It became apparent very soon that the change in the imperative singular of strong verbs with vowel gradation could not be examined with the help of "traditional" corpora (Mair, 2012) Although some of them are comparably large (e.g. DeReKo[1]) and contain spoken language (e.g. corpora in the DGD database[2]), where linguistic change usually starts out before it finds its way into written language, none of these corpora yielded enough tokens of the target imperative singular forms for a systematic (let alone statistical) analysis. Two reasons for the rarity of the imperative singular can be found in the Duden grammar (Dudenredaktion, 2009, 548-550): its use is tied to the condition that speakers are on familiar terms (use of the informal second person singular pronoun *du*), and there are several other constructions used instead of the imperative to express requests or commands, such as indicative, modal and infinitive constructions.

Pre-existing large web-based corpora also have drawbacks. Most of them do not provide meta information about the authors of texts, a circumstance which has rather obvious reasons, given the wealth of data in the corpora, and which could be accepted. More serious for the study of a recent language change is the fact that no information is available about when the texts in these corpora were produced, as is the case, for example, in the deWaC (Baroni et al., 2008).

2 The Walkthrough Corpus

Instead of consulting traditional or existing web-based corpora, a corpus was specifically compiled for the present investigation. It consists of a web-specific text type, viz. walkthroughs, which contains a high number of instances of imperative singular forms. In addition, the website which was crawled contains very recent language material and the majority of texts on it have a timestamp, so that the development of imperative formation can be tracked.

2.1 Texts

Walkthroughs are guides for video games, i.e. computer, console and internet games, which help gamers complete a game successfully. They include step-by-step instructions, lists of achievements and items, cheats and other tips. Like in official strategy guides (usually in print), which are commissioned by the game publishers, their main focus is on a precise rendering of the game's content. In contrast to the former, these online guides are written by gamers and the texts are subject to a minimal amount of proofreading or revision. The conditions of their production are therefore very close to natural language.

Perhaps most importantly for the present investigation, the fact that members of the gaming community write walkthroughs for other members provides for an increased use of the imperative singular, e.g. *nimm den Gegenstand* 'take the item', *erstich den Feind* 'stab the enemy' etc., which is otherwise only rarely attested in corpora of German, spoken or written. A pilot search on the web revealed several candidate websites for the corpus

[1] http://www.ids-mannheim.de/cosmas2/

[2] http://dgd.ids-mannheim.de/

compilation, only one of which provided some of the required meta information (also see section 3 below):

The website *spieletipps.de*[3] exists since 2001 (in the present form). It was crawled in 2013; hence, the corpus covers a time span of 12 years. It is one of the main gaming websites in Germany, on which complete walkthroughs, individual cheats and tips and forums are available for the majority of existing platforms (including retro ones like Atari consoles). The final walkthrough corpus compiled from the website comprises approximately 7 m. tokens or word forms.

2.2 Crawler

A webcrawler (Java) was tailored to the website in order to download all walkthrough texts, cheats etc. Each text was stored in one line of a csv file. Available meta information about texts and authors were similarly stored in separate csv files. When queries were entered in the search interface (2.4), the data from these files were reunited through an inverted index.

2.3 Annotation

All texts contained in the corpus were then tagged for their part of speech using the TreeTagger (Schmid, 1995). This should enable the search for imperative forms of verbs (POS-tag VVIMP in STTS) and thereby reduce the number of word-level queries (however, see 3.1). The annotated versions of all texts were similarly stored in a csv file.

2.4 Interface

A simple search interface was created, comparable to those of popular web search engines. It allowed word-level, e.g. *gib*, and POS-level queries, e.g. *vvimp2geben* 'imperative forms of give'. It outputs csv files with one row for each query hit and columns for the query, sentence context and meta information.

3 Challenges

Challenges arose during the compilation of the corpus, the search for imperative singular forms in it and the annotation of the data for additional variables. One of these can be attributed to the researcher (3.1); others are specific to the website (3.2 and 3.3), the corpus (3.4), the walkthrough genre (3.5 and 3.6) or the search interface (3.7).

3.1 Non-computational linguists

Linguists who want to investigate a potential language change-in-progress might find themselves in a situation when the phenomenon in question is not or only rarely attested in "traditional" corpora. Even though they might be able to perform a pilot search using one of the major web search engines, many (if not to say) most linguists do not possess the necessary programming skills for the compilation of a corpus of web data.

Since this was the case in the present paper, the compilation of the corpus itself was left in the hands of a computer scientist. However, the latter needs to be carefully instructed by the researcher in order that the final product yields the required results. Thus, the linguist should have a precise idea not only of which data and meta data are available during crawling (see 3.2 and 3.3) but also of how annotation for additional variables may be partly automatised by the use of an appropriate interface (see 3.7).

3.2 Meta information about corpus texts

Although the corpus compiler in the present case was instructed to retrieve each text on the website along with all available meta information, he can only include data in the corpus which is provided by the website (creators). A crucial piece of information for an investigation of language change in general, and perhaps ongoing change in particular, is the point in time when a linguistic utterance was produced.

Unfortunately, the original timestamp of posts on the website used for the present analysis was not given. However, in contrast to corpora such as the deWaC, which do not provide a date, either, two dates could be retrieved from the present website: i) when a member had registered, and ii) when the game to which the entry referred was released in Germany (or the earliest universal release, if no German version exists). The timestamp was extrapolated as the more recent of these two dates: a member cannot post a walkthrough or other tip for an existing game on the website before being registered, and even as a registered member, he/she cannot post a walkthrough or anything similar about a game which has not been released yet. The format of the timestamps was

[3] http://spieletipps.de

mixed; they were therefore reduced to only the posting year.

It turned out later that on the profiles of members, their last postings were listed with the original posting date. The comparison of the original and extrapolated timestamps for the instances in the final dataset revealed that 53.6% of the extrapolated posting years were correct and 22.0% could be replaced by the years listed on the member profiles. For the remaining 24.4% of observations, only the extrapolated dates were available (due to the author's active membership). Separate statistical analyses performed on the full dataset and a reduced dataset without these observations showed that extrapolation did not have an effect on the results of the investigation.

3.3 Meta information about authors

In times of heated discussion about data protection, it is easily understood that members of a website or forum wish to remain anonymous. On the website used for the present investigation, members can theoretically provide personal information such as their full name, age and residence on their profile page, and they can select which of these data to share with the public. The crawler could only include meta information about the authors of texts in the corpus which was visible on their public profile page. Therefore, in the final dataset, which was used for the statistical analysis, only 21.3% of the instances had an annotation for the author's age, 13.4% for gender (based on members' first names), and 6.8% for their residence. Analyses of the influence of sociolinguistic factors on the change in imperative singular formation of the strong verbs with vowel gradation were thus based on such small samples that they identified trends, but the results are not generalisable.

3.4 POS tagging

As mentioned before, the corpus search interface allowed word-level and POS-level queries. Unsurprisingly, the analogical e-stem imperative variants of verbs were incorrectly tagged as finite forms or as proper names; therefore, instead of using the POS-level query, these forms (e.g. *geb*, *gebe*) had to be searched on word-level for every individual verb. Perhaps more interestingly, even though the i-stem imperative is the established variant, only lower case instances of, for example, *gib* were recognised correctly, whereas capitalised *Gib* was often incorrectly tagged as a noun or proper name and therefore not returned by the POS-level query.

The available options in the corpus search interface were thus sufficient to extract hits on the word (and POS) level, and the immediate sentence context in the output files provided enough information to distinguish finite forms of a verb from genuine imperative hits (1 vs. 2) and imperative forms of simplex verbs from those of particle verbs (2 vs. 3):

> (1) "Ich hab das Spiel bereits durchgespielt und **gebe** hier mal die Gegner bekannt"
> (nds/fluch-karibik-3/2620615)
>
> (2) "**Gebe** ihm das Glas und die Spirale"
> (pc/clever-smart/2742515)
>
> (3) "**Gebe** die ersten beiden Buchstaben ein"
> (snes/nba-jam/310511)

However, at least the distinction between examples (1) and (2) would have been largely performed by the TreeTagger if it had been trained accordingly, which would have reduced the amount of manual POS-tag correction. In similar investigations of variation or recent language change, it may be worth adapting the TreeTagger or supplying manually tagged training material with instances of the target construction or form before tagging the actual corpus texts.

3.5 Authorship confirmation

Although all imperative singular instances in the dataset were annotated for the member of the website who had contributed the text in which they occurred, it had to be ascertained that all of these instances were indeed produced by the specified author. For a number of reasons, the authors quote from inside the game whose walkthrough they are writing. Some of these quotes are readily recognisable as such from the use of quotation marks or their occurrence in tables of so-called "achievements", such as "Stiehl 30 Fahrzeuge" 'Steal 30 vehicles' (Gangstar Miami Vindication for iPhone). Other quoted imperative singular forms occurred in running text without any indication of being borrowed. The consultation of "Let's Play" videos[4] proved an efficient way of exposing the unmarked in-game imperative uses. In

[4] Youtube - http://youtube.com

these videos, gamers tape their computer or console screen while playing a particular game and comment on how (missions or chapters in) the game can be completed successfully. Thus, any in-game commands which were quoted in the walkthrough appear on the player's screen in the video and can be discarded from the dataset.

3.6 Skewed frequency data

While walkthroughs have the advantage of being practically the only text type to contain a very high number of instances of the imperative singular, their special topic presented another challenge. One of the aims of the present project was to test whether the Conserving Effect of high token frequency in analogical change is also found in the recent change in imperative singular formation of German strong verbs with vowel gradation. To this end, instances in the dataset should be annotated for the verb's token frequency in German.

Unfortunately, the plots of video games are very different from everyday life in the real world; therefore, token frequencies of words in walkthroughs are necessarily skewed. For example, avatars in first- and third-person shooters and a number of role-playing games do not eat, but *essen* 'to eat' is a strong verb with vowel gradation in German, hence one of the target verbs. If the token frequencies for this and other target verbs had been taken from the walkthrough corpus, the results of the analysis would have been skewed as well. In order to avoid this, verb token frequencies needed to be extracted from reference corpora (DeReWo[5]; Projekt Wortschatz Universität Leipzig[6]), frequency dictionaries (Jones and Tschirner, 2006; Ruoff, 1990), and frequency counts provided in a dictionary of German (Duden online[7]).

3.7 Annotation for persistence

A close reading of some of the texts in the corpus revealed that the specific form of the imperative may in part depend on the preceding context. Benedikt Szmrecsanyi explained that language users are "creatures of habit" and tend to reuse words or patterns whenever possible (2005; 2006). This "persistence" strategy may be at work in the formation of the imperative singular of the strong verbs with vowel gradation as well.

Imperative singular forms of German verbs can usually occur as a suffixed or unsuffixed variant: *red*!/ *rede!* 'talk!', *renn*!/ *renne!* 'run!', *steh*!/ *stehe!* 'stand!'. Similarly, the analogical e-stem variants of strong verbs with vowel gradation can occur with or without the suffix -e: *nehm*!/ *nehme!* 'take!'; however, the i-stem variant is never suffixed: *nimm*! It seemed that the authors of the walkthroughs developed a "routine", so that when they had used the suffixed variants of one or several consecutive verbs, e.g. *laufe ... gehe ... verlasse*, they wished to add a suffix to the following imperative singular form as well. If this next verb was a strong verb with vowel gradation, the author has no choice but to use the suffixed analogical e-stem variant because a suffixed i-stem imperative singular variant of these verbs does not exist. Examples (4) and (5) illustrate this persistence effect of suffixed and unsuffixed previous imperatives.

(4) -e → -e
"2. Stell**e** deine Gäste einander vor und verkuppl**e** sie.
3. G**e**b**e** deinen Gästen genügend zu trinken, indem..."
(ps2/playboy-mansion/2260012)

(5) -ø → -ø
"Nach der Cutszene, geh**-ø** zu Junes und geh**-ø** in die TV-Welt. Sobald du drinnen bist spr**i**ch**-ø** mit Rise um den letzten Boss zu suchen."
(ps2/persona-4/3379622)

In order to test this hypothesis, all instances of imperative singular forms of strong verbs with vowel gradation in the dataset need to be annotated for the form of imperative singular occurrences in their preceding context. As the interface which was created for the walkthrough corpus does not incorporate context queries, all imperative singular forms preceding the target forms in the dataset were searched and annotated manually.

4 Joys

The compilation of the walkthrough corpus and the search for and annotation of relevant instances of the target construction presented many challenges. Results which would otherwise not have been obtained, however, by far outweigh the costs of manual labour. Not only did the corpus study

[5] http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html

[6] http://wortschatz.uni-leipzig.de/

[7] http://www.duden.de/woerterbuch/

reveal frequency and persistence effects on imperative singular formation of the strong verbs with vowel gradation, but these results also served as input for a subsequent experimental study.

4.1 Results of the corpus study

After removing all false hits, the final dataset comprised 1939 observations of imperative singular forms of strong verbs with vowel gradation, i.e. instances of the established i-stem variant and the suffixed and unsuffixed analogical e-stem variants. Mixed-effects regression models were fitted on the dataset in order to determine which of the annotated predictor variables had an influence on stem vowel choice and suffixation of the imperative singular forms.

As expected, verb token frequency has a significant effect on stem vowel choice: imperative singular forms of lower frequency verbs show a high probability of occurring with the analogical e-stem, while higher frequency verbs retain the established i-stem. The Conserving Effect of high token frequency in analogical change is thus confirmed for morphological change-in-progress in German (see Figure 1).

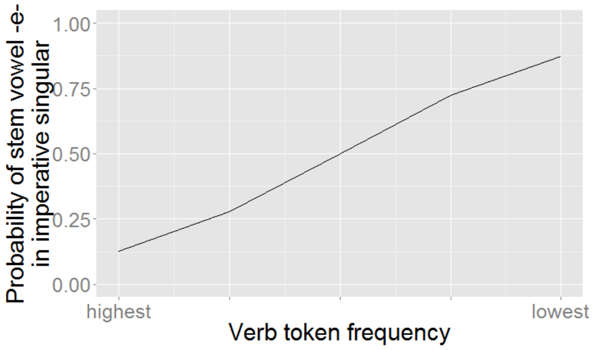

Figure 1: Conserving Effect of verb token frequency

The analysis also affirmed that the suffixation of the imperative forms (and thereby indirectly stem vowel choice) is significantly influenced by the occurrence of suffixed or unsuffixed imperative singular forms prior to the target imperatives: Imperative singular forms of strong verbs with vowel gradation show a high probability of being suffixed (e.g. *nehme*) when they are preceded by other suffixed imperative singular forms. Unsuffixed variants of the imperative singular of these verbs (e.g. *nimm*, *nehm*) occur more often after other unsuffixed imperative singular forms (see examples 4 and 5). This effect is reinforced when the previously occurring verb itself is a strong verb with vowel gradation, e.g. *nehme* following *gebe* or *nimm* following *gib*.

4.2 Experimental Study

The Conserving Effect of high token frequency is generally explained on the basis of “entrenchment” (Langacker, 1987, 59): through repeated use, the imperative forms of higher frequency verbs have stronger mental representations than those of lower frequency verbs. Therefore, the forms of higher frequency verbs are more quickly retrieved from memory than the forms of lower frequency verbs. The longer they take to retrieve, the higher is the probability that the speaker forms the imperative in analogy to the weak verb paradigm. An instance of recent language change, such as the example of imperative singular formation examined in the present paper, is an excellent test case for the validity of this assumption.

In the experiment conducted as part of the current project, participants’ reactions to the established i-stem and analogical e-stem imperative singular variants, presented in verbs of different token frequency, were measured. Once it was known from the corpus study that, in addition to the predictor verb token frequency, the presence of suffixed or unsuffixed imperative singular forms prior to the target imperative has a significant influence on the formation of the imperative singular of strong verbs with vowel gradation, this potentially disturbing persistence effect could be eliminated in the experiment and stimulus design. Furthermore, the corpus study showed up trends with regard to the influence of dialect on imperative formation (3.3) which inspired the inclusion of participant groups from different dialect areas in order to test this notion more systematically than was possible in the corpus study itself. Finally, sentences adapted from walkthrough texts can accommodate a large number of verbs from diverse semantic fields without appearing too absurd to the participants. Thus, the corpus texts served yet another purpose.

5 Suggestions for future research

The text type used in the present investigation was identified through the coincidence that the author relied on walkthroughs in order to complete several video games and was therefore aware of the high number of imperative singular forms contained in texts of this kind. In other studies, suitable web-specific genres/ text types may be identified by performing pilot searches on the web or in existing large web corpora and inspecting whether instances of a target construction predominate in a particular text type or web register. The situation might be further improved by attempts at recognising and classifying as many web registers as possible and identifying linguistic patterns associated with them (cf. Egbert and Biber, 2013; Biber et al., 2015).

As concerns the compilation of a corpus for a linguistic study, this task should preferably be delegated to a person who has experience with working with a linguistic corpus or is familiar with the kinds of questions linguists wish to answer with the help of corpora. In the current study, the presence of an "intermediary" or "translator", i.e. a linguist with extended IT knowledge, proved helpful while the research assistant was instructed on how to compile the walkthrough corpus (cf. 3.1). At the same time, the intermediary could answer the author's questions about how the corpus and its query interface are created.

However, even the best assistant (and intermediary) has to rely on the needs and demands which the employing researcher expresses, who in turn has to know the website(s) and features of the specific text type very well. If the website *spieletipps.de* would have been more thoroughly inspected before corpus compilation, the time stamps for walkthroughs could have been extracted primarily from the member profile pages; only if they were not available there, the programme would have to resort to the release date of the video game and the member registration date as a proxy (3.2). Similarly, as explained above, the persistence variables were manually annotated (3.7), i.e. the verb class and suffixation of imperative singular forms in the preceding context and the textual distance to the target imperative form were searched and counted by hand. Slight changes to the crawler could have reduced the amount of manual labour in both annotation steps. As the analysis of corpus data is at least as time-consuming as the compilation of a corpus, researchers might be tempted to push compilation forward before knowing the included sources well enough. The present investigation illustrates clearly that the manual effort which can be avoided outweighs the costs of a thorough inspection of potential corpus texts, e.g. particular websites.

The analysis of sociolinguistic patterns of variation according to authors' age, gender and location only revealed trends in the present corpus study. In such cases, conducting additional studies, for example a psycholinguistic experiment (4.2) is an effective way of consolidating or falsifying these trends.

6 Conclusion

Even though the challenges of using web-based corpora for analysing recent language change seem to outweigh the joys in the present contribution, this is largely due to the fact that the former have been more elaborately discussed in order to serve as advice for researchers of similar phenomena in future. Some of the manual labour explained above might be increased in web-specific genres of the walkthrough kind and cannot be avoided completely, such as extracting frequency data from reference corpora or other sources. Other drawbacks of the present corpus have been avoided in the compilation of large existing web-based corpora: for example, the DECOW (German web corpus by COW; Schäfer, 2015) is annotated for meta information like the "last modified" date. And yet others may be avoided by tailoring the corpus compilation process to the specific object of study, e.g. adjusting or training the TreeTagger on a target construction.

Nevertheless, it cannot be stressed too often that the walkthrough corpus offered data without which work on the present project would not have been possible. A bit of manual labour (after hours) was rewarded with many invaluable insights from the corpus analysis. Not only do these results explain the present stage of the change-in-progress in imperative formation of strong verbs with vowel gradation, but they also find their repercussions in the design of a subsequent experimental study.

Acknowledgements

Work on the present PhD project (including the compilation of the walkthrough corpus and payment for participants in the experiment) was funded by the *Deutsche Forschungsgemeinschaft* through the Research Training Group GRK DFG 1624 "Frequency Effects in Language".

I would like to thank three anonymous WAC-X reviewers for their helpful comments.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2008. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43:209–226. [deWaC].

Douglas Biber, Jesse Egbert, and Mark Davies. 2015. Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, 10:11–45.

Joan Bybee and Sandra Thompson. 1997. Three frequency effects in syntax. In Matthew L. Juge and Jeri L. Moxley, editors, *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure*, pages 378–388.

Dudenredaktion, editor. 2009. *Duden - Die Grammatik: unentbehrlich für richtiges Deutsch*. Dudenverlag, Mannheim and Wien and Zürich.

Jesse Egbert and Douglas Biber. 2013. Developing a User-based Method of Web Register Classification. In *Proceedings of the 8th Web as Corpus Workshop*, pages 16–23.

Joan B. Hooper. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William Christie, editor, *Current progress in historical linguistics*, pages 95–105. North Holland, Amsterdam.

Randall L. Jones and Erwin Tschirner. 2006. *A Frequency Dictionary of German: Core Vocabulary for Learners*. Routledge, London.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar: vol. 1: Theoretical Prerequisites*. Stanford University Press, Stanford.

Christian Mair. 2012. From opportunistic to systematic use of the web as corpus. In Terttu Nevalainen and Elisabeth Traugott, editors, *The Oxford Handbook of the History of English*, pages 245–255. Oxford University Press, New York.

Arno Ruoff. 1990. *Häufigkeitswörterbuch gesprochener Sprache*. Niemeyer, Tübingen.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora*, pages 28–34. [DECOW14].

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 1–9. [TreeTagger].

Thomas Schmidt. 2014. Gesprächskorpora und Gesprächsdatenbanken am Beispiel von FOLK und DGD. *Gesprächsforschung*, 15:196–233. [DGD].

Benedikt Szmrecsanyi. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, 1(1):113–150.

Benedikt Szmrecsanyi. 2006. *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Trends in Linguistics. Studies and Monographs. De Gruyter, Berlin and New York.