15

Bamberger Beiträge zur Linguistik

A Blend of MaLT

Selected Contributions from the Methods and Linguistic Theories Symposium 2015

Hanna Christ, Daniel Klenovšak, Lukas Sönning,
Valentin Werner (eds.)

University
of Bamberg
Press

15 Bamberger Beiträge zur Linguistik

Bamberger Beiträge zur Linguistik

hg. von Geoffrey Haig, Thomas Becker (†), Martin Haase,
Sebastian Kempgen, Manfred Krug und
Patrizia Noel Aziz Hanna

Band 15

University
of Bamberg
Press
2016

A Blend of MaLT

Selected contributions from the Methods and Linguistic Theories Symposium 2015

Hanna Christ, Daniel Klenovšak, Lukas Sönning,
Valentin Werner (eds.)

University
of Bamberg
Press
2016

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.d-nb.de/ abrufbar.

Herstellung und Druck: Digital Print Group, Nürnberg
Umschlaggestaltung: University of Bamberg Press, Anna Hitthaler

http://www.uni-bamberg.de/ubp/

ISSN: 2190-3298
ISBN: 978-3-86309-457-7 (Druckausgabe)
eISBN: 978-3-86309-458-4 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus4-482257

Table of contents

Part 1: Methodological reflections and impulses

Part 2: Bringing methods and linguistic theories together

Introduction

Hanna Christ, Daniel Klenovšak, Lukas Sönning and Valentin Werner
University of Bamberg

This volume presents selected contributions to MaLT 2015, the first *Methods and Linguistic Theories Symposium,* which was organized by the *Bamberg Graduate School of Linguistics* and hosted at the University of Bamberg from November 27th–28th 2015. The focus of the event was on bringing together research methods and theory, and – we were glad to realize – it struck a timely theme. On the one hand, this is due to the fact that over the past few decades, linguistics as a field has undergone a major transformation, evolving from a predominantly descriptive to an increasingly empirical discipline. Even a cursory glance at current linguistic journal volumes clearly reveals this shift. At the beginning of the 21st century, it is fair to acknowledge that research on language relies heavily on empirical and quantitative evidence.

Many would agree that linguistic theorizing has benefited from this transformation. Once largely dominated by introspective methodology, we now have at our disposal additional tools for inductive reasoning, and, perhaps more importantly, for assessing the adequacy of established models and theories. In recent decades, our field has gained new insights by *directly* turning to the object of knowledge – that is, language and how it is used and processed by humans. As such, data may suggest new routes toward understanding and question familiar paths. Upon confrontation with quantitative evidence, our formalized state of knowledge may have to be refined or rethought.

In the course of this transformation, methodological know-how has become one of the key qualifications for researchers, especially young academics at the beginning of their careers. New skills are required to find a way through the quantitative maze in the literature and to choose a sensible approach for the particular phenomenon on one’s desk. Learning from data is both an art and a science. In a field where widespread use of empirical methodologies is a relatively recent development, it may at times be difficult to acquire the relevant (statistical and other) literacy.

However, it is a fact that linguistics curricula at the tertiary level commonly lack training in empirical methods, so that young researchers more often than not have to resort either to external offers (for instance workshops organized by professional organizations) or become self-taught (provided their institutional libraries contain adequate resources). Further, the empirical turn in linguistics has gone hand in hand with a considerable diversification of research methods. While this diversity has come to be seen as a strength of linguistics as a field, the plethora of procedures may puzzle even the seasoned researcher. Still, ignoring methodological developments is not an option if meaningful linguistic research is to be conducted. In the light of the current vibrancy of the interplay between research methods and theory building, the aim of MaLT was twofold:

(i) to provide a forum for researchers to meet peers from other branches of linguistics;
(ii) to provide a venue to look beyond specific disciplinary boundaries and draw inspiration from neighboring fields.

The emphasis on cross-disciplinary exchange offered researchers the opportunity to expand their repertoire of theoretical approaches and methods within and beyond those typically adopted in their subfields.

The conference was thus conceptualized as an ensemble of talks and practical workshops, which offered hands-on advice in two broad fields currently taking center stage in the empirical study of linguistic structures: corpus linguistics and experimental linguistics. In the former area, Samantha Laporte (University of Louvain) introduced the hows, whats and whys of corpus linguistics in her workshop *What corpora can do for you: An introduction to corpus methods and corpus tools*. Quantitative methods for handling corpus data were discussed in a practical *Introduction to statistics for corpus linguistics* by Stefan Evert (University of Erlangen-Nürnberg). Two workshops focused on experimental linguistics. Franziska Günther (Ludwig-Maximilians-University of München) discussed the fundamentals of experimental work in *How to collect (and combine) linguistic and behavioral data: A practical workshop on experiments in linguistics*. Participants also had the opportunity to delve deeper into the state-of-the-art toolbox of psycholinguists, with Franziska Hartung's (MPI Nijmegen) workshop *Experimental methods in discourse processing*. The general program was rounded off by two topical plenary talks by Alexander Ziem (University of Düsseldorf) titled *From discourses to corpora: Cognitive approaches to (lexical) meaning-making* and by Martin Hilpert (University of Neuchâtel) on *How to blend MALT: Bringing methods and linguistic theory together*. We owe heavily to the latter for inspiration for the title of this volume and would like to express our gratitude to all workshop conveners and plenary speakers for the time and effort invested.

With close to 100 participants from more than 10 countries around the globe, MaLT can be considered a great success. With the program being aimed at early-career researchers, one main concern of the organizers was to grant participation in the conference and the workshops free of charge. It thus goes without saying that MaLT 2015 would not have been possible without generous financial support. In particular, we would like to thank the German Academic Exchange Service (DAAD) for supporting MaLT through its IPID4all scheme. We also received considerable funding from the University of Bamberg (FNK) and the alumni represented by the Universitätsbund Bamberg e.V.

As an event such as MaLT is very much a collective effort, we would like to extend our gratitude to all those who supported the symposium in various ways, first and foremost to Marion Hacke and Simone Treiber from the *Trimberg Research Academy* (TRAc) of the University of Bamberg. Further we would like to thank Geoffrey Haig and Hans-Ingo Radatz as speakers of the *Bamberg Graduate School of Linguistics* for their input, all those involved chairing the individual sessions (Hanna Budig, Romina Buttafoco, and Ole Schützler), and the student helpers (Carolin Cholotta and Katharina Scheiner), who ensured a smooth running of the whole event.

Further, as regards the preparation of this volume, we would like to say thank you to all authors for their efforts in preparing and revising their manuscripts, and for their feedback as internal reviewers for other papers. In addition, Romina Buttafoco, Jiří Milička, Jochen Podelo, Ole Schützler, and Fabian Vetter acted as external referees and provided helpful suggestions to improve the overall quality of the individual papers.

That said, in this book we are proud to present a selection of the contributions. They can be seen as the essence of what MaLT was about, and nicely illustrate the range of topics covered as well as the various concerns and approaches that featured during the event.

The first part is predominantly oriented toward crucial aspects relating to linguistic methodology in terms of data types, collection, presentation and analysis.

Alexander Ziem opens the volume with a paper titled *From discourses to corpora: (lexical) meaning-making as a challenge for cognitive semantics*, which discusses the use of corpus-linguistic tools in cognitive-linguistic discourse analysis. Navigating within a cognitive-functional framework, the analyses are grounded in the assumption that linguistic meaning emerges from language use. The primary object of study is what Ziem refers to as "U-relevant knowledge": language users' cumulative and collective knowledge about linguistic signs. The paper exemplifies empirical procedures for the investigation of how U-relevant knowledge is shaped in discourse. Methods at different levels of analysis are illustrated with data from discourses on sociopolitical and economic crises in Germany over the past 40 years. First, Ziem demonstrates the use of exploratory lexicometric techniques, which not only serve to provide a birds-eye perspective on lexical patterns across discourses, but also yield insights for hypothesis generation. To add substance to the abstract numbers provided by multifactorial analyses, word frequency distributions are compared to identify lexis that is specific to a particular discourse, or shared across two or more discourses. Ziem then shows how analytical categories borrowed from functional-cognitive grammar allow the researcher to "zoom in" further to uncover different conceptualizations of shared lexical items. A frame-based analysis of lexical meaning may thus detect more fine-grained differences in the way concepts are used and essentially shaped in discourse. Throughout the paper, Ziem's central concern is to illustrate how discourse analysis can benefit from the use of corpus-linguistic methods.

The two following chapters deal with methodological aspects of sociolinguistic fieldwork. Adina Staicov sets the scene with her paper *Methodologies in sociolinguistic fieldwork*, in which she provides practical advice on a wide range of issues involved in collecting sociolinguistic data in the field. The author discusses essential steps in planning and carrying out field research and gives valuable insights based on her own experience, which she gathered in research projects on different varieties of English, including the Fiji islands in the South Pacific, British Asians in London and the San Francisco Chinatown community. Her advice carefully balances technical, cultural, and personal reflection. Throughout her contribution, she stresses the importance of knowledge and awareness of the target community, which may have critical implications for the researcher's conduct. Staicov's contribution is valuable for anybody planning to enter the field, as her advice and experience reports sensitize the reader to potential challenges along the way.

In another contribution relating to linguistic fieldwork, this time for the study of variation in a non-native variety of English, Sofia Rüdiger introduces an innovative approach to the elicitation of conversational material. In *Cuppa coffee? Challenges and opportunities of compiling a conversational English corpus in an Expanding Circle setting* she first contrasts written and spoken linguistic data and discusses both why spoken data – despite their often-cited primacy – are understudied, and why data allegedly "spoken" often underlie certain constraints (for instance in terms of a formal setting during sociolinguistic interviews) that preclude an analysis as "conversational" and "naturalistic". Rüdiger continues to argue that many of these constraints can be avoided if a truly informal interview setting is established, and she proposes what she labels the "cuppa coffee method", where interviewer and interviewee engage in mutual exchange over a cup of coffee (used as "social lubricant") in a public space. Like in traditional interview approaches, parts of the conversation are recorded and thus can be subject to linguistic analysis. Rüdiger also points out potential drawbacks of the "cuppa coffee method" such as increased transcribing time or potential recording quality issues. That the method developed by her is not merely an intellectual game is shown in the last part of her chapter where she details how her approach resulted in the compilation of the *Spoken Korean English Corpus* (SPOKE) used to analyze naturalistic speech.

Collecting linguistic data via online experiments is a mixed blessing, as is shown by Jana Häussler and Tom Juzek. In their contribution *Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgment task*, the authors point out that recruiting participants via crowdsourcing platforms like *Amazon's Mechanical Turk* is a cheap and easy way of collecting quantitative data. However, focusing on the observations in their acceptability judgment tasks, the authors also cast doubt on the reliability of these data, and show that participants often exhibit non-cooperative behavior in the sense of merely "clicking their way through". Thus, participants are negligent of actually performing the task, which potentially influences the quality of the overall results in a negative way. Through careful examination of their methodology, Häussler and Juzek provide the reader with ways of detecting such behavior based on response times. They further discuss some ideas on how to circumvent and discourage unaccommodating ratings like the implementation of booby trap items and the tracking of response times in order to keep the data as "clean" as possible.

Moving on to visualization methods in quantitative research, the last contribution of part one presents a relatively unfamiliar type of display – *The dot plot: A graphical tool for data analysis and presentation*. In his paper, Lukas Sönning introduces and illustrates the dot plot and argues for its routine usage in quantitative research. Based on principles of graph construction and empirical evidence from research into visual perception, advantages of dot plots over other commonly used chart types such as the bar chart are demonstrated. The paper outlines design options and extensions and illustrates the application of this chart type in linguistic data analysis, including examples from corpus linguistics and meta-analysis. Sönning also reflects on its limitations and provides Microsoft Excel spreadsheet templates for the production of dot plots.

The papers in the second part aim to show how varying methodological approaches or changing methodological parameters can affect the interpretation of results, which may yield different implications for linguistic theory building.

In the field of word formation, Chiara Naccarato illustrates how the notoriously vague concept of productivity can be assessed using quantitative diachronic data. Her paper *A corpus-based quantitative approach to the study of morphological productivity in diachrony: The case of samo-compounds in Russian* investigates the changing productivity of the Russian prefixoid *samo-* from 1700 to the present day. In her concise analysis, the author applies Baayen's measure of "potential productivity" and discusses its major disadvantage: unreliability due to artifacts when it is applied to corpora of different sizes, yielding a result of supposedly decreasing productivity. This problem is overcome through the use of a Large Number of Rare Events model estimating the expected number of types and *hapax legomena* with *samo-*. Thus, Naccarato is able to demonstrate the increase in productivity of *samo-* over time. She goes on to analyze the productivity of different compound patterns with *samo-* in detail, confirming the frequently assumed interrelation of lexicalization and productivity: highly lexicalized words of high frequency form a small group and are based on less productive patterns (*samolet* 'aircraft', *samogon* 'moonshine') whereas productive patterns produce a large number of low-frequency items.

The volume is completed by an empirical assessment of the German pluralization system by Eugen Zaretzky and Benjamin P. Lange, in which they argue that *No matter how hard we try: Still no default plural marker in nonce nouns in Modern High German*. In their paper, the authors analyze how various intralinguistic factors, such as grammatical gender, word-final phonemes, plural markers of the rhyming real words, unusual orthography, final-obstruent devoicing, etc., condition the choice of plural allomorphs in nonce words, such as *Pind* → *Pinder*, in a sample of German native speakers. Comparing their findings to an earlier study with the same test items, their main methodological aim is to show that (i) the sample size, (ii) the type of regression, and (iii) particularly the study design, mainly in terms of task types (plausibility rating vs. production) used, may markedly influence the overall results. Based on their quantitative evidence, they identify a number of weaknesses of earlier approaches and eventually suggest that, instead of dual-route models, which have been advocated in previous studies, single-route models best account for the distribution of plural markers.

Part 1

Methodological reflections and impulses

From discourses to corpora: (lexical) meaning-making as a challenge for cognitive semantics[1]

Alexander Ziem
University of Düsseldorf

Many discourse-semantic studies focus on the role of lexical units (e.g. key words, focal words, stigmatizing words, buzz words, etc.) as "carriers" of ideological framing, emphasizing multiple ways of coining and framing word meaning within public discourse, including mass media. In this field of research, one of the most exciting issues is to identify and describe strategies of semantic coining in public language use. In contemporary political discourse in Germany, for example, buzz words such as *Heuschrecke* ('locust', metaphorical for financial investor) and *Herdprämie* ('bonus to remain at the cooking stove') do not only provide access to rich lexical meanings but, more importantly, also to an entire discourse. How can patterns of semantic coining be identified in corpora? How can discursive processes of meaning-making be investigated? Taking these questions as starting points, this paper pursues three aims. First, it introduces a cognitive- and discourse-linguistic framework as well as a set of linguistic categories, including frames, semantic roles, respectively frame elements, and (argument structure) constructions relevant for scrutinizing ways of shaping concepts in discourse. Second, it reports on corpus technologies, particularly the software *Lexico3*, useful for quantitatively comparing related discourses in terms of their vocabulary. Finally, it summarizes the results of a corpus study on lexical meaning-making in discourses on "crises".

1. Coining meaning in discourse: Introductory remarks

To what extent does language use mark social, more specifically political reality? With which linguistic means is (collective, but also individual)

[1] The empirical part of this paper draws on results achieved in a DFG-funded research project on linguistic construals of "crises" in Germany since 1973. I would like to thank the project collaborators Kristin Kuck, David Römer, Ronny Scholz, and Martin Wengeler for helping build up the text corpus and carrying out joint corpus analyses. I am also deeply indebted to Paul Sarazin, who provided native speaker support for preparing this paper. Finally, I would like to thank three anonymous reviewers for their very helpful comments. The usual disclaimer applies.

"knowledge" about social events and circumstances created, rationalized, and enforced by discourse actors? How, by virtue of language, is the impression generated that with reference to social, public-political issues something specific is the case – what then constitutes our "knowledge" of this "object"? Starting from interrelated questions of this type, discourse-linguistic analyses have the task of elucidating how collective knowledge is created, negotiated, and disseminated in social discourses. In contrast to critical discourse analysis, it is not their primary concern to conduct studies from a socially critical standpoint, and, in so doing, occupy one's own position in order to perform evaluations from that position. Instead, the discourse-linguistic approach relates to the concept of an "archaeology of knowledge" (loosely based on Foucault 1973) or to a linguistic epistemology (see Section 2.1).

Among those "social facts" worthy of being investigated are sociopolitical and economic crises, whose half-life is intimately interrelated to the public discourse. As soon as the mass media stop debating "social facts", they disappear from (public) awareness. And, conversely, as soon as they slide into the center of daily reporting, crises turn into "facts", which become ever more concrete with each additional day they are topicalized in the media. The fact that socially severe "crises" – like the "financial crisis" of 2008/09 or the current "government debt crisis" – emerge and entrench within mass media discourses and, as a result, mark our *knowledge by description* (Warnke 2009), makes a discourse-linguistic analysis of the linguistic procedures employed relevant beyond the boundaries of the discipline.

Which concepts are negotiated in discourses on crises? To what extent are various discourses on crises similar or different? Which linguistic indicators can be used to reconstruct central concepts with respect to their "coinedness" in discourse? Which "knowledge" about crises prevails, which actors determine this "knowledge"? Which political measures are legitimated, and how are they legitimated? Which methods are appropriate in order to be able to answer such questions? In the following, the discourse-semantic approach chosen for that study will be presented, taking examples from these discourses on "crises". First, Section 2 describes the methodological framework in which the current approach is situated.

Subsequently, Section 3 exemplifies to what extent quantitative lexical analyses allow for an initial classification of the data. This illustrative part of the article draws on results obtained in the aforementioned research project and comprehensively documented in respective project publications.[2]

2. Investigating U-relevant knowledge

2.1. Discourse linguistics as linguistic epistemology

The leading assumption of discourse linguistics is that shared linguistic knowledge of a language community can be determined with the help of corpus-linguistic research (of a quantitative and/or qualitative nature). Such endeavors need to take into consideration

> that discourses not only comprise the superficial level of lexical meanings of the linguistic signs used in the discourse, but also want to capture the semantic prerequisites, implications, and possibility conditions that are characteristic of individual statements. (translated from Busse & Teubert 1994: 23)

To the extent that discourse-semantic studies are at the service of linguistic epistemology (Busse 2008), the decisive analytic aim is to capture the knowledge of a language community that is relevant to understand a linguistic expression as comprehensively and extensively as possible (Ziem 2014: 150–172). In this view, linguistic meanings are epiphenomena in that they emerge from conceptualization processes (Fauconnier & Turner 2002). Accordingly, linguistic structures are at the same time understood as the results of communicative usage of linguistic expressions and as a sedimented stock of knowledge which language users possess to comprehend and to describe. This type of knowledge will hereafter be referred to as U-relevant knowledge ("Understanding-relevant knowledge"; Ziem 2014: 133).

[2] See Wengeler & Ziem (2014) and Ziem (2017, in press), as well as the collected volume Wengeler & Ziem (2013), in methodological respects also Ziem (2013a), among others. The DFG-funded research project on linguistic construals of social and economic-political crises in Germany since 1973 was directed by Martin Wengeler and myself, supported by David Römer, Ronny Scholz, and Kristin Kuck.

The methodological starting point is the assumption that meanings of linguistic signs are not to be found in the signs themselves. Rather, linguistic signs allude to knowledge, which the receiver has to update on the basis of co-text, context, and his or her prior knowledge each time. On the basis of minimal linguistic input, such as a single lexeme, highly complex knowledge relations become available, are ordered by relevance criteria and correspondingly "contextualized" (in the sense of Busse 2007). Following the concept of a linguistic epistemology, meanings can only be adequately described when the communicative situation is taken into account (Busse 1987: 272), which is why the smallest unit of a discourse analysis is a usage event. Discourses form the culturally marked and historically variable frame within which their communicative sense becomes possible. Correspondingly, discourse semantics is guided by the "collective knowledge of a discourse community in a given epoch with regard to the topical area chosen as the object of investigation or the semantic field or discourse formation" (translated from Busse 1987: 267).

With its orientation to collective knowledge, this concept of discourse can hardly be separated from the concept of U-relevant knowledge. U-relevant knowledge becomes effective during comprehension processes, and can be differentiated in three ways (cf. for example Busse 1991: 149–150): with a view to (a) levels of knowledge, referring to paradigmatic and syntagmatic aspects of the organization of signs, for instance, (b) modes of knowledge, that is, truth values, which can be attributed to or withheld from a proposition uttered, and which can vary between the poles of "taken to be certain" and "taken to be false", and (c) types of knowledge (such as linguistic knowledge, knowledge of social forms of action/interaction, everyday practical knowledge of actions, individual knowledge of experiences, etc.). As a critical comparison with a psycholinguistically motivated classification of knowledge types shows (Ziem 2010a), there is much to indicate that with these knowledge types one is not only dealing with heuristic items; rather, their cognitive relevance has been observed on numerous occasions in empirical studies (cf. Graesser et al. 1997, among others).

2.2. How can U-relevant knowledge be investigated?

U-relevant knowledge is not only exchanged within a linguistic community, but also coined in discourse. How can these coinages be approached empirically? Busse & Teubert (1994: 14) made the groundbreaking suggestion of defining discourses for research-practical purposes as “virtual text corpora, whose constitution is determined in the broadest sense by content (or semantic) criteria.” The elucidations of these criteria suggest (and this is how the definition has been received within discourse research) that a corpus under investigation should in particular contain thematically aligned texts.

To what extent, however, can *virtual* corpora form the object of discourse-semantic analysis? It is not possible to conduct empirical research on virtual corpora; rather, they serve as a standard. Discourses are *virtual* corpora insofar as they have empirically hardly traceable dimensions: Thematic corpora used as the basis for certain, discourse-linguistic research remain principally incomplete, as – in spite of huge electronic resources – it will never be possible to capture all thematically pertinent texts exhaustively. It follows that in practice it is possible to study only a partial amount of the potentially relevant texts. Empirical discourse research is thus inevitably oriented towards “mere” discourse fragments. A *real* text corpus correspondingly forms a subset of the respective discourse. Valid conclusions can therefore only be drawn about the population of discourse represented by a compiled corpus. Therefore, Busse & Teubert (1996: 14–16) have correctly pointed out that the linguistic object of discourse only comes into being in the course of building the corpus. Similar to the linguistic units “word”, “sentence” and “text”, a discourse is not simply present, but rather is the result of theory-led empirical observations. The constitution of a specific text corpus is thus a research task, which always has to be geared to what is practically realizable.

A constitutive semantic criterion of corpus building is, according to Busse & Teubert (1994: 14), a common communication and epistemic context of the constituent texts. Most of the other criteria can be subsumed thereunder. At first, this suggestion is motivated by the research-practical requirement of empirically investigating discourses from a linguistic perspective. Another reason is that meaning-making in discourse always takes place within a common thematic context. The findings of cognitive-linguistic experiments support this claim. For example, pieces of information about overarching thematic context – which in newspaper articles is shown in headings (Brône & Coulson 2010) or which can be presumed by knowledge of the text (see Vu et al. 2000) – have demonstrable priming effects on the semantic conceptualization of newly introduced discourse referents and ambiguous terms. In other words, thematic relations co-determine linguistic meanings. Based on these findings, analyses of thematic corpora have shown that even within a very short period of time salient knowledge facets become so deeply entrenched that language users presuppose them as shared background knowledge (Ziem 2014: 289–314). In relation to the example analysis in Section 3, I will deal with such sedimentations in data from discourses on crises.

2.3. Areas of discourse-linguistic investigations

It is the merit of Warnke & Spitzmüller (2008) to have developed a procedural-practical model for linguistic discourse analyses. The model takes into account methodological and empirical considerations in the literature, and provides an answer to the discourse-historically relevant question of "how one should properly take language in discourse de facto into account". The model is conceived as "a practical operationalization, which corresponds to the methodological presuppositions of discourse linguistics" (translated from Warnke & Spitzmüller 2008: 23). Without being able to elucidate the model in detail at this point, Table 1 provides an overview of the relevant analytical categories.

The model suggested has the advantage of clarifying the numerous methodological approaches and related accompanying challenges for linguistic discourse analysis. However, it has the disadvantage that one cannot study all aspects named in specific analyses. Moreover, some of the classifications summarized in Table 1 are controversial, for example the question of why "historicity" does not equally concern the intratextual level beyond the transtextual level. It is also unclear to what extent key words and stigma words, for instance, really represent intratextual phenomena when they are distinguished by the fact that they provide relations implicitly spanning across texts, and indeed become key and stigma words by their appearance in various texts but similar co-texts. This is also true for metaphors.

Table 1. Discourse-linguistic levels and categories ("DIMEAN") following Warnke & Spitzmüller (2008: 49)

Transtextual level	
Discourse-oriented	Schemata (frames/scripts), basic discourse-semantic figures, topoi, social symbolism, indexical orders, historicity, ideologies/mentalities, general social and political debates
Actors	
Interaction roles (author, anticipated addressees)	
Discourse positions (social stratification/power verticality status, discourse communities, ideology, brokers, voice)	
Mediality (medium, forms and areas of communication, text patterns)	
Action (constitution of issue, connection of issue, evaluation of issue)	
Intratextual level	
Text-oriented	Visual text structure, layout/design, typography, text-image-relations, materiality/medium Macro-structure: text topic/ meso-structure: topics in parts of the text, lexical fields, metaphor fields, lexical lines of opposition, development of topics, text strategies/text functions, text genres
Proposition-oriented	Micro-structure: propositions (syntax; rhetorical figures; metaphor lexemes; social, expressive, deontic meaning; presuppositions; implicatures; speech acts)
Word-oriented	Multi-word units/single-word-units/morphology/word formation (key words, stigma words, names, ad-hoc-formations)
Sound-oriented	Phonology/phonetics (conversation-analytical units of investigation)

Methodologically, an added value of discourse analysis (in comparison to other approaches limiting their area of investigation to the level of the word, sentence, or text) lies instead in the fact that *principally every* analytical category can be placed on the transtextual level. Every category can come into use, when it is a matter of the *exemplariness* of speech acts, presuppositions, rhetorical figures, text-image relations, typography, etc. To that extent a separation of trans- and intratextual level, as undertaken by Warnke & Spitzmüller (2008), appears redundant from a discourse analytical perspective.

The model proposed by Warnke & Spitzmüller as displayed in Table 1 comprises three levels with a multitude of sub-categories, into which specific units of analysis and description are classified. Regardless of the concerns expressed, the analytical categories used in the following (specifically in the example analyses in Section 3) largely correspond to those used in this model. In particular, it is vital to mention that it is possible to identify and replicably describe meaning-making in discourse – the subject matter of the following – within a series of discourse events by means of those linguistic analysis categories located on the transtextual level of the model. Frames, basic discourse-semantic figures (Ziem 2014: 339–349) and argumentation schemata ("topoi"; Wengeler 2003) have proved to be particularly helpful tools (for an overview see Wengeler & Ziem 2014). Furthermore, in the analysis it is also possible to include conceptual metaphors/metaphor fields as well as key words on the transtextual level (Kuck & Römer 2012). Yet, in the discourse-analytic framework ("DIMEAN") proposed by Warnke & Spitzmüller (cf. Table 1) they are classified into the intratextual level; still, they are particularly of interest as discourse phenomena, that is, as elements of the transtextual level.

3. Meaning-making in discourses on crises

In the following, I present results from diachronic studies, using examples from discourses on crises in the Federal Republic of Germany. The focus lies on the presentation and application of text-statistical methods (for a more comprehensive overview see Ziem 2017, in press). The results achieved can serve as the starting point for qualitative studies, as, for example, for discourse-semantic analyses of metaphoric (Kuck & Römer 2012; Ziem 2009, 2014: 315–378) and lexical meaning-making (Scholz & Ziem 2013; Ziem 2013b). For reasons of space, I will not present ways of deepening analyses in detail.

The object of analysis are discourses on crises in Germany since the so-called "oil crisis" in 1973/1974. In addition to (a) the "oil crisis" itself, it concerns (b) those economic and sociopolitical events significant for the so-called "geistig-moralische Wende" ('intellectual and moral turn') in 1982, (c) the discussions of "reform" and "Wirtschaftsstandort Deutschland" ('economic location Germany') of the 1990s, for which the label "Arbeitsmarktkrise" ('labor market crisis') became established in 1997, (d) the debates on the future of the welfare state within the framework of "Agenda 2010" culminating in the measures of 2003, as well as (e) the so-called "Finanzkrise" ('financial crisis') in 2008/2009. The study is based on a text corpus encompassing approximately 11,000 texts, which were selected by means of a systematic key word search from five German newspapers of record – *Bild, Frankfurter Allgemeine Zeitung, Süddeutsche Zeitung, Der Spiegel* and *Die Zeit*. The main criterion for corpus generation was that of the mutual communication context (Busse & Teubert 1994: 23), that is, the explicit focus in terms of content on one of the five social and economic-political crises since 1973. All the relevant texts were digitized and, where necessary, transformed into machine-readable documents, as well as provided with a meta-data head, so that it was possible to flexibly build sub-corpora according to research interest (see for example Scholz & Wengeler 2012; Ziem et al. 2013). The corpus-management program INGWER, which was developed in collaboration with semtracks, enabled systematic control of the corpus (for details see Ziem et al. 2013).

An important aim of the discourse-linguistic study was to contribute to the analysis of events marking Germany in terms of its history of mentalities. More specifically, motivated by the concept of a linguistic epistemology (e.g. Busse 2008; Ziem 2013a), we strive for working out linguistic mechanisms with which social crises have been co-created and marked in the public media since 1973. Four levels and categories of investigation were taken into account in particular. For the present purposes, I will only go into the first one; I will, however, come to mention key words, insofar as these can be derived from the text-statistical research results:

- *Multifactorial analysis, common and specific lexis:* Discourses on crises demonstrate differences and similarities at the lexical level. While common lexis points to discursive similarity, specific lexis shows discursive divergence (Scholz & Mattissek 2014).
- *Key words:* Key words indicate salient epistemic elements of a discourse. Moreover, they often make disputable topics visible. This is, for example, the case when lexical meanings in a given context are meta-linguistically discussed among discourse participants (Stötzel & Wengeler 1995).
- *(Conceptual) metaphors:* Metaphors make abstract concepts linguistically accessible in that they are described in terms of something that is familiar, as, for example, the conceptualization of a crisis as a natural disaster (see for example Böke 1996, 1997; Ziem 2014: 322–332, 371–376).
- *Argumentation schemata:* Lexical units like key words and metaphors are used in texts as argumentative tools; they appear as parts of arguments. If they appear in patterns, they can consolidate to topoi (cf. Wengeler 2003).

Methodologically, the systematic evaluation of the corpus proceeded as follows: first, quantitative corpus-linguistic procedures – such as the tools provided by the software *Lexico3* (http://lexi-co.com/ressources/ manuel-3.41.pdf) – were used. They allow for initial sorting, in particular by means of discourse-comparative analyses of the lexis of a crisis. Second, subcorpora were annotated semantically to study linguistic coinages of key concepts. These annotations mainly followed frame-semantic prerequisites, although frames did not only function as coding schemata, but also as interpretative-analytical instruments which allowed to draw inferences about cognitive fixations of conceptual knowledge. Finally, a hermeneutic text-analytical approach to metaphors and argumentation schemata was necessary: with differing search criteria (see, for instance, Kuck & Römer 2012: 74) selected articles were read in their entirety, metaphor areas and argumentation topoi annotated in the database in order to afterwards be able to analyze their occurrence in different time periods, individual newspapers and various actors.

The purpose of the initial quantitative approach to the text corpus is thus to look through the discourse-constituting data as comprehensively as possible, in order to then conduct an inductive analysis in toto. Scharloth et al. (2013: 348) describe this step as follows:

> Instead of checking a hypothesis with previously defined categories of analysis, all patterns in a corpus are calculated, which result from the application of previously set algorithms. These patterns are categorized afterwards. In so doing, pieces of evidence often come into focus which are either contrary to the previously existing expectations or in the ideal scenario such evidence forms the basis for new hypotheses which suggest the formation of new interpretative linguistic categories of analysis. (translated from Scharloth et al. 2013: 348)

With regard to lexis, such a "corpus-driven" (Tognini-Bonelli 2001: 84–101) approach primarily consists of calculating the lexical recurrences, without prior formulation of expected patterns, such as the relevance of individual lexical items in discourse. The focus is on questions such as the following:

- What constitutes the specific lexis of a discourse?
- To what extent does the lexis of a discourse overlap with the lexical inventory of another discourse?
- Do certain word fields structure a discourse to such an extent that they can be taken as characteristic of it?
- To what extent do individual lexical units dominate a discourse?
- Can specific or idiosyncratic discourse characteristics be inferred?

These key questions already indicate that reliable results from studies of lexis are to be expected, in particular with discourse-comparative analyses. With recourse to lexicometrical methods this will be illustrated in the following sections. Lexicometry is an excellent tool for discourse-semantic analyses of lexis because it enables researchers to gain hypotheses about dominating patterns of language use. Particularly helpful are multifactorial analysis (Section 3.1) and the calculation of the specific and mutual lexis of (sub-)discourses on crises (Section 3.2).

3.1. Multifactorial analysis

Multifactorial analysis is a statistical tool which facilitates the calculation of frequency relations of non-lemmatized word forms between (sub-)-discourses as well as its diagrammatical illustration (Lebart et al. 1998: 45). In contrast to many current procedures, lexicometry forgoes any lemmatization of word forms because it is seen as an interpretative intervention into the raw material of the corpus. Even though lemmatizations prove necessary for many studies, they are not always unproblematic following discourse-linguistic premises, because, inter alia, grammatical forms also fulfill relevant, ideological functions in discourse (see for example Hart 2014), from which one should not abstract away.

A multifactorial analysis allows to determine and clarify commonalities and differences in the lexis of several (sub-)discourses. This happens on the quantitative basis of a comparison of the relative frequency of occurrence of all word forms in a corpus with the relative frequency of occurrence of all word forms in a reference corpus (or several reference corpora). More specifically, the aim is to compare the lexis used in various discourses, on the basis of digital text corpora. The corpora are to be composed previously according to the criterion of their thematic relevance and further research guiding principles (in the sense of Busse & Teubert 1994).

With respect to discourses on crises, each corpus, or sub-discourse (here: the "oil crisis" in 1973, the "intellectual and moral turn" in 1982, the "labor market crisis" in 1997, the "Agenda 2010" in 2003, and the "financial crisis" in 2008/2009), covers a sociopolitical or economic crisis. Multifactorial analysis groups word forms that occur in a sub-discourse with a high frequency, into a "factor."

> With factor analysis a large number of different variables – in the present case linguistic properties – is reduced to a small number of derived variables, the factors. Thereby each factor represents a summary or generalization of properties, which co-occur at high frequency. In our

> study word forms which, for example, co-occur in part of a corpus [i.e. a sub-discourse as explicated above, AZ] with a high frequency, are grouped together as a factor. (translated from Scholz & Ziem 2013: 159)

The factors can be presented in a coordinate system, from which the degree of similarity between the lexis of sub-discourses can be read off on the basis of the distance between the factors of the individual sub-discourses.

Figure 1 clearly shows that a larger similarity exists between the lexis of “Agenda 2010” and the “labor market crisis” than between all other sub-discourses. In turn, the “oil crisis” takes on a special role; the largest differences are found between its lexis and that of all the other crises. Furthermore, the lexis in the sub-discourse on the “financial crisis” demonstrates differences from the lexis of all the other sub-discourses – with one exception: the “oil crisis”. Finally, it is noticeable that the lexis of the discourse on the “party political transition”, on the “labor market crisis” and around “Agenda 2010” share more commonalities with each other than with the “oil crisis” or with the “financial crisis”. This suggests a greater discursive proximity of the former and, in turn, a special status of the latter.

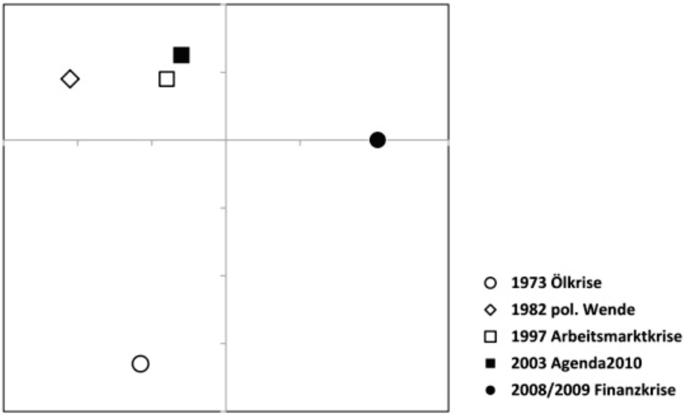


Figure 1. Multifactorial analysis of the entire corpus (Scholz & Ziem 2013: Section 3.1)

The result of the multifactorial analysis summarized in Figure 1 emanates solely from the quantitative comparison of the word forms contained within a corpus. Multifactorial analysis admittedly does not allow for an analytical grasp of lexical meaning-making; however, the purely quantitative comparison of the lexis used in a particular discourses uncovers general similarities and differences at this level of description. This corpus-linguistic grasp of the lexis of political language use therefore fulfills the purpose of sharpening one's view of notable properties, which can be followed up in further steps of analysis.

One discourse-historically striking point of note relates to the finding that the "oil crisis" and "financial crisis" strongly diverge from the other discourses on crises at the level of lexis. Conversely, the "labor market crisis" and "Agenda 2010" display particularly noticeable commonalities at the lexical level. A lexicometric calculation of the respective specific and mutual lexis lends itself well in order to get to the bottom of the features. In the following, I shall present these quantitative corpus-linguistic approaches to lexis using the example of the comparison of the discourses on the "labor market crisis" and "Agenda 2010" in more detail.

3.2. Lexicometric analyses of the lexis of discourses

Beyond multifactorial analysis, lexicometrical tools allow calculations of the common lexis of (sub-)discourses which is based on word forms with the same relative frequency in the (sub-)discourses. The common lexis indicates to what extent (sub-)discourses are similar with respect to their lexis. This not only uncovers common tendencies in terms of choice of word forms, but also exposes common thematic foci, especially if one takes into account the mutual nominal lexis.

Moreover, statistical tests for the calculation of probability distributions (for instance by means of a chi-square-test or hyper-geometric distributions as in *Lexico3*) can be used to determine the specific lexis, that is, word forms that are overrepresented in a (sub-)discourse. The determination of the specific lexis is based on significance tests. More precisely, for each word form it is determined whether frequency differences between corpora are statistically significant. If that is the case, then its relative frequency in the corpus is not coincidental.

Table 2. Discourse-historical comparison of the specific lexis (rank 1–25; S > 50) (Ziem et al. 2013: 162)

Labor market crisis	Agenda 2010
Deutschland ('Germany')	*SPD* (Sozialdemokratische Partei Deutschlands)
Arbeitslosigkeit ('unemployment')	*Schröder* (former German Chancellor)
Mark (former German currency)	*Deutschland* ('Germany')
Arbeit ('work')	*Kanzler* ('Chancellor')
Arbeitsplätze ('jobs')	*Gewerkschaften* ('trade unions')
Kohl (former German Chancellor)	*Eichel* (former German finance minister)
Arbeitsmarkt ('labor market')	*Union* ('union')
Steuerreform ('tax reform')	*Reformen* ('reforms')
Waigel (former German finance minister)	*Arbeit* ('work')
Spiegel (German weekly magazine)	*Clement* (former German economics minister)
Reform ('reform')	*Agenda* ('agenda')
Lafontaine (former German finance minister)	*Reform* ('reform')
Währungsunion ('currency union')	*2010*
Sozialen ('social')	*Kommission* ('committee')
Beschäftigung ('employment')	*Wachstum* ('growth')
Arbeitslosen ('unemployed')	*Gerhard* (first name of the former German Chancellor Schröder)
DM (former German currency)	*2004*
Schaffen ('create')	*Schröders*
Arbeitgeber ('employers')	*2003*
1997	*Sozialhilfe* ('social welfare')
1996	*Kündigungsschutz* ('dismissal protection')
Leistungen ('payments')	*Steuerreform* ('tax reform')
Sozialhilfe ('social welfare')	*Stoiber* (former prime minister of Bavaria)
Senkung ('reduction')	*Müntefering* (former German finance minister)

There are various statistical procedures for the calculation of significance. First, hyper-geometric distributions (Lebart et al. 1998: 129) underlie the software *Lexico3*, developed for lexicometric purposes. Applied to relative frequencies of word forms in a corpus or a sub-discourse, it is thereby possible to determine the amount of those word forms that are quantitatively overrepresented in a corpus. They constitute the specific lexis of this corpus.

To illustrate, the comparison of the specific lexis of all five crises, that is, all five corpora, provides an overview of significantly frequent word forms within a sub-discourse in comparison to all other sub-discourses. Hence, it is well possible that, for instance, one word form is overrepresented in the specific lexis of two or even more sub-discourses at the same time, since the significance results from the rare occurrence of the word form in the other sub-discourses. One such interesting case is that of the word forms *Arbeit* ('work'), *Reform* ('reform'), *Steuerreform* ('tax reform'), *Sozialhilfe* ('welfare') in the sub-discourses on the "Agenda 2010" and "labor market crisis"; in both, these word forms belong to the specific lexis.

Based on the findings summarized in Table 2, we may speculate on why certain (groups of) word tokens are strongly overrepresented within a discourse. Likewise, it is possible to formulate hypotheses about what causes a particular word token to belong to the specific lexis of two (or several) sub-corpora at the same time. I will briefly illustrate below how lexicometrically obtained quantitative results could be used in terms of research practice and guidelines.

Because the word forms *Reform, Steuerreform, Sozialhilfe,* and *Arbeit* belong to the specific lexis of both the "labor market crisis" and the "Agenda 2010" crisis, that is, they are strongly overrepresented in these corpora in comparison to the lexis of the other crises, it seems reasonable to assume that these word forms belong to the common lexis of both crises. Is this really the case? Is it a valid conclusion for all four word tokens in the same way? Which word forms are particularly strongly overrepresented? In order to be able to give answers to these questions, one viable alternative is to systematically determine the mutual nominal lexis. Ordered by descending frequency, Table 3 gives an overview of the ten most frequent nominal word tokens, which are equally overrepresented in the corpus on the "labor market crisis" and on "Agenda 2010".

To the extent that frequency of occurrence of a linguistic unit indicates its degree of relevance in a discourse, *Sozialhilfe* is a good candidate for a controversial concept in public discourse. At least in quantitative terms it plays a fundamental role in both sub-discourses. *Sozialhilfe* is a key word whose conceptual coinage in discourse requires fine-grained semantic investigations.

Table 3. Absolute frequency of nominal word tokens of the mutual lexis in the domain of social policy (rank 1–10; S > 50)

	Absolute frequency	
Token	**Labor market crisis**	**Agenda 2010**
Sozialhilfe ('welfare')	290	421
Sozialstaat ('welfare state')	185	245
Lohnnebenkosten ('payroll taxes')	161	235
Arbeitgebern ('employers')	81	96
Sozialabgaben ('social security contributions')	75	84
Sozialpolitik ('welfare policy')	74	132
Abgaben ('expenditures')	74	100
Wohlstand ('prosperity')	74	91
Sicherungssysteme ('security system')	70	121
Arbeitsmarktes ('(of the) labor market')	64	107

3.3. Key word analysis

The key word *Sozialhilfe* is very strongly overrepresented in social-political discourses on "Agenda 2010" and the "labor market crisis", and thus belongs to the common lexis of these discourses. However, this does not inevitably mean that the key word *Sozialhilfe* is used synonymously in both discourses. Rather, a detailed semantic analysis is necessary in order to demonstrate possible individual, discourse-specific semantic coinages. To this end, frame semantics has proven a suitable analytical instrument (Kalwa 2010; Storjohann & Schröter 2011; Ziem 2014; see also Ziem 2013a for a summary).

The frame-semantic corpus analysis is based on a three-step procedure (for more details see Scholz & Ziem 2013; Ziem 2014: 349–361): (a) extraction of the sentences to be annotated, here: those sentences in which *Sozialhilfe* appears; (b) annotation of the frame elements realized; (c) qualitative evaluation of the annotations and interpretation of the results obtained. A look at the extracted sentences shows that the frame elements are realized in particular in possessive constructions (e.g. *Kürzung der Sozialhilfe* 'reduction of welfare') and nominal phrases with an attributive adjective (e.g. *bisherige Sozialhilfe* 'welfare payments to date'); they were thus taken into consideration during the annotation process. We set out with the 14 frame elements, of the [ASSISTANCE]-frame evoked by the lexical unit *Sozialhilfe*, as annotation categories.[3] It was not possible, however, to adequately capture numerous predicates purely on the basis of these 14 frame elements. Among them was a multitude of predicates which lead to a conceptualization of *Sozialhilfe* as an affected object (e.g. *Kürzung* 'cutback', *Beschneidung der Sozialhilfe* 'cutting of welfare payments'). An additional semantic role was therefore defined, concerning qualitative descriptions, which allows *Sozialhilfe* to become an affected object.

After due analysis of the annotated linguistic realizations of the frame elements, it appears that in the context of "Agenda 2010", *Sozialhilfe* is described seven times more often with respect to its inherent properties, in particular its amount (as in *Niveau* 'level'/*Höhe der Sozialhilfe* 'amount of welfare payment', *niedrige Sozialhilfe* 'low welfare payments') than in the context of "labor market crisis". The relevant predicates relate to the frame element MANNER; it thus plays a major role in the "Agenda 2010". Furthermore, within the discourse on the "labor market crisis" *Sozialhilfe* becomes an affected object more than twice as often; typical for this pattern are possessive constructions such as *Einsparung der Sozialhilfe* ('saving of social welfare'), *Absenkung der Sozialhilfe* ('lowering of social welfare'), *Reform der Sozialhilfe* ('reform of social welfare'). Such a conspicuously frequent topicalization of the looming decrease in social-security payments within the framework of the "labor market crisis" points to a more pronounced social-political awareness of problems in the media. In the context of "Agenda 2010"

[3] See https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=luIndex; the respective frame elements are BENEFITED PARTY, GOAL, FOCAL ENTITY, HELPER, DEGREE, DURATION, EXPLANATION, FREQUENCY, INSTRUMENT, MANNER, MEANS, PURPOSE, PLACE, and TIME.

this scarcely played a role any longer insofar as merely inherent-qualitative properties of welfare are named in the discourse.

As becomes clear from this example, frames as documented in *FrameNet* (https://framenet.icsi.berkeley.edu) can be used for discourse-semantic analytical purposes. This is possible in that frame elements serve as annotation categories which takes as a prerequisite that the respective relevant frame is captured in the *FrameNet* database. If that is the case, it can nevertheless turn out that the respective pertinent frame elements are not sufficiently differentiated for the annotation. As a follow-up, there exists the option of defining and using additional frame elements for the analysis based on the data to be annotated.

3.4. Embedding structures of key words

Word meanings and lexical meaning-makings in general are at least partly the result of co- and contextual embedding of the target expression. An abstraction of these embedding structures inevitably leads to an incomplete analysis. Without being able to present these studies in detail here (for an overview see Ziem 2017, in press), taking the example of political key words, I would at least like to illustrate to what extent superordinate embedding structures, alongside local ones, should also be systematically included in analyses of lexical meaning-making.

Why is it problematic in lexical-semantic analyses to disregard the syntactic as well as superordinate structures (such as text genre, medium)? The most important reason is arguably the following: The conceptualization of the terminological content of a linguistic expression varies depending on (a) the semantic role realized (cf. e.g. *They are fighting [the crisis]*$_{\text{Affected\_object.}}$ vs. *They are evaluating [the crisis]*$_{\text{Theme}}$) and (b) the syntactic function (e.g. *[The crisis]*$_{\text{Subject}}$ *is shaking Germany* vs. *Germany is overcoming [the crisis]*$_{\text{Direct\_object}}$). In each of the examples given, *Krise* (‘crisis’) is the key word, and the syntactic functions and semantic roles are annotated as indices (the latter drawing on von Polenz 2008: 167–174). Beyond the embedding of a key word in such argument-structure constructions, a complete analysis has to consider further embedding structures, which potentially mark word meanings in discourse. Among them are the following:

- *Collocations and multi-word units*: As mentioned above, the significantly frequent occurrence of one word with another word (e.g. Teubert 2002; Storjohann & Schröter 2011) or several other words ("n-grams", Bubenhofer in press; "collocations", see Steyer 2013) can give an indication of salient facets of meaning within a discourse.
- *Topos*: Often the lexical content of a word stands in close connection to its recurrent argumentative absorption in certain topoi (in the sense of Wengeler 2003). If this is the case, the argumentative potential of a (key) word sediments to constitute part of its lexical meaning (Ziem 2014b).
- *Text type/genre*: The meaning of an expression can be co-determined by the text type or genre within which the expression is embedded.
- *Medium*: Systematic variations in meaning can be effects of the medium with which a linguistic expression is realized (as an integral component of a text). Here, *medium* is understood as a broader concept, which should enable one to differentiate both between media of the same type (e.g. *Frankfurter Allgemeine Zeitung* vs. *Bild*-Zeitung) and between categorically distinct media (for instance chat vs. face-to-face communication).
- *Discourse*: In many cases, meaning-making occurs within a thematically aligned communication context, for instance the metaphor for financial investors as *Heuschrecken* ('locusts') in the debate on capitalism (Ziem 2014: 315–379). If the coinage is strong, that is, the degree of conventionalization is high, then it can also endure beyond the discourse (as was, and still is, the case with *Heuschrecke*).

The influence of local and superordinate embedding structures on the meaning-making of lexical units can be identified systematically based on corpus investigations. Taking the example of the key word *Krise* ('crisis'), I have tried to show this with a view to both (a) local embedding structure (argument-structures) and (b) newspaper medium (cf. Ziem 2013c). Embedding structures are hence determinants to be taken into account in the process of lexical meaning-making.

4. Concluding remarks

The discourse-linguistic phenomenon of lexical meaning-making was the central subject matter of the present article. The starting point was the observation that ideological positions and attitudes can sediment in word meanings which are deemed to be emergent results of gradual coinages within a communication community. Such sedimentation processes take place under the conditions of complex communicative settings which shape lexical meanings of the words used. Inter alia, the linguistic context, the discourse participants, the (mass-)medium and the thematic-content relation in which the linguistic expressions under investigation are used, belong to the parameters of complex communicative settings. An adequate semantic analysis has to account for these parameters as comprehensively as possible.

Taking the example of discourses on crises in Germany, I have tried to illustrate in which form a large text corpus can first be "sorted" and structured by corpus-linguistic means. The summary presentation of research results should help to show that a discourse-semantic approach provides an appropriate apparatus to trace and suitably describe linguistic coinages in social knowledge production. It is hence capable of making a contribution to a linguistically reflected clarification of mechanisms of linguistic constructions and constitutions of social "facts" or "issues".

The study of meaning-making pertains to the core tasks of the analytical approach demonstrated. For the analysis of meaning-making the relevance of corpora can hardly be overstated. Such a research perspective first and foremost aims at making statements, beyond observed individual findings, about *regularities* in the discourse (linguistic patterns), that is, beyond *typical* linguistic units with *typical* properties, which have gradually formed in the course of language use. Thus, corpus-linguistic analyses of lexis in political language use require both (a) a corpus-driven approach that helps to formulate first hypotheses about lexical meaning-making by means of machine procedures and (b) fine-grained corpus-based semantic analyses that take account of contextual embedding structures. In the present contribution, for reasons of space, I have discussed this model only very briefly. Elsewhere, however, I have demonstrated to what extent a frame-semantic research approach can take on these tasks (Ziem 2014).

An overarching aim of the article was to provide an overview of current corpus-linguistic procedures that can do justice to these demands. The account limited itself to a selection of lexicometric tools for analyzing the lexis of discourses. Specifically, the focus lay on (a) quantitative analyses of lexis by means of multifactorial analysis, (b) collection of the mutual lexis, that is, those word forms quantitatively equally overrepresented in various corpora, and (c) the collection of the specific lexis, that is, those word forms which are significantly frequent. These corpus-linguistic approaches prepare semantic analyses, but they are not to be seen as analyses of lexical meanings. Instead, their task consists of identifying dominant and salient word forms in a corpus compared to reference corpora.

In the realm of language use in mass media, the lexicometric methodology chosen is at the service of "linguistic epistemology" (Busse 2008): It serves as a useful means for the empirical research of U-relevant knowledge and its specific coinage in public language use. The increasing diversity of corpus-linguistic opportunities for studying language use with respect to regularities opens a broad horizon for future research.

References

Böke, Karin. 1996. Überlegungen zu einer Metaphernanalyse im Dienste einer "parzellierten Sprachgeschichtsschreibung". In Karin Böke, Matthias Jung & Martin Wengeler (eds.), *Öffentlicher Sprachgebrauch: Praktische, theoretische und historische Perspektiven*, 431–452. Opladen: Westdeutscher Verlag.

Böke, Karin. 1997. Die "Invasion" aus den "Armenhäusern Europas": Metaphern im Einwanderungsdiskurs. In Matthias Jung & Martin Wengeler (eds.), *Die Sprache des Migrationsdiskurses: Das Reden über "Ausländer" in Medien, Politik und Alltag*, 163–192. Opladen: Westdeutscher Verlag.

Brône, Geert & Seana Coulson. 2010. Processing deliberate ambiguity in headlines: Double grounding. *Discourse Processes* 47(3). 212–236.

Bubenhofer, Noah. In press. Kollokationen, n-Gramme, Mehrworteinheiten. In Kersten Sven Roth, Martin Wengeler & Alexander Ziem (eds.), *Handbuch Sprache in Politik und Gesellschaft*. Berlin: Mouton de Gruyter.

Busse, Dietrich. 1987. *Historische Semantik: Analyse eines Programms*. Stuttgart: Klett-Cotta.

Busse, Dietrich. 1991. *Textinterpretation: Sprachtheoretische Grundlagen einer explikativen Semantik*. Opladen: Westdeutscher Verlag.

Busse, Dietrich. 2007. Diskurslinguistik als Kontextualisierung: Methodische Kriterien. Sprachwissenschaftliche Überlegungen zur Analyse gesellschaftlichen Wissens. In Ingo H. Warnke (ed.), *Diskurslinguistik nach Foucault: Theorien und Gegenstände*, 81–105. Berlin: Mouton de Gruyter.

Busse, Dietrich. 2008. Linguistische Epistemologie: Zur Konvergenz von kognitiver und kulturwissenschaftlicher Semantik am Beispiel von Begriffsgeschichte, Diskursanalyse und Frame-Semantik. In Heidrun Kämper & Ludwig M. Eichinger (eds.), *Sprache – Kognition – Kultur: Sprache zwischen mentaler Struktur und kultureller Prägung*, 73–114. Berlin: Mouton de Gruyter.

Busse, Dietrich & Wolfgang Teubert. 1994. Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In Dietrich Busse, Fritz Hermanns & Wolfgang Teubert (eds.), *Begriffsgeschichte und Diskursgeschichte: Methodenfragen und Forschungsergebnisse der historischen Semantik*, 10–28. Opladen: Westdeutscher Verlag.

Fauconnier, Gilles & Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Foucault, Michel. 1973. *Die Archäologie des Wissens*. Frankfurt am Main: Suhrkamp.

Stötzel, Georg & Martin Wengeler (eds.). 1995. *Kontroverse Begriffe: Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*. Berlin: Mouton de Gruyter.

Graesser, Arthur C., Keith K. Millis & Rolf A. Zwaan. 1997. Discourse comprehension. *Annual Review of Psychology* 46. 163–189.

Hart, Christopher. 2014. *Discourse, Grammar and Ideology: Functional and Cognitive Perspectives.* London: Bloomsbury Academic.

Kalwa, Nina. 2010. Islam-Konzepte im Kölner Moscheebaudiskurs: Eine framesemantische Analyse zum Islambegriff. *Aptum* 1. 55–75.

Kuck, Kristin & David Römer. 2012. Metaphern und Argumentationsmuster im Mediendiskurs zur ‚Finanzkrise'. In Kathrin Lämmle, Anja Peltzer & Andreas Wagenknecht (eds.), *Krise, Cash und Kommunikation – Die Finanzkrise in den Medien*, 71–95. Konstanz: UVK.

Lebart, Ludovic, André Salem & Lisette Berry. 1998. *Exploring textual data*. Dordrecht: Springer.

Mattissek Annika & Ronny Scholz. 2014. Der Hochschulreformdiskurs: Eine Diskursanalyse mit Hilfe lexikometrischer Verfahren. In Johannes Angermüller, Martin Nonhoff, Eva Herschinger, Felicitas Macgilchrist, Martin Reisigl, Juliette Wedl, Daniel Wrana & Alexander Ziem (eds.), *Diskursforschung. Ein interdisziplinäres Handbuch. Band 2: Methoden und Praxis der Diskursanalyse. Perspektiven auf Hochschulreform-diskurse,* 86–112. Bielefeld: Transcript.

Polenz, Peter von. 2008. *Deutsche Satzsemantik: Grundbegriffe des Zwischen-den-Zeilen-Lesens*. Berlin: Mouton de Gruyter.

Scharloth, Joachim, David Eugster & Noah Bubenhofer. 2013. Das Wuchern der Rhizome: Linguistische Diskursanalyse und Data-driven Turn. In Dietrich Busse & Wolfgang Teubert (eds.), *Linguistische Diskursanalyse: Neue Perspektiven,* 345–380. Wiesbaden: Springer.

Scholz, Ronny & Alexander Ziem. 2013. Lexikometrie meets FrameNet: das Vokabular der "Arbeitsmarktkrise" und der "Agenda 2010" im Wandel. In Martin Wengeler & Alexander Ziem (eds.), *Sprachliche Konstruktionen von Krisen: Interdisziplinäre Perspektiven auf ein fortwährend aktuelles Phänomen,* 155–184. Bremen: Hempen.

Scholz, Ronny & Martin Wengeler. 2012. "Steuern runter macht Deutschland munter" und "Kriegen die Pleitebanker auch noch einen Bonus?": Zwei Wirtschaftskrisen in Bild. *Osnabrücker Beiträge zur Sprachtheorie* 81. 155–176.

Steyer, Kathrin. 2013. *Usuelle Wortverbindungen: Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

Storjohann, Petra & Melanie Schröter. 2011. Die Ordnung des öffentlichen Diskurses der Wirtschaftskrise und die (Un-)Ordnung des Ausgeblendeten. *Aptum* 1- 32–53.

Teubert, Wolfgang. 2002. Die Bedeutung von Globalisierung. In Oswald Panagl & Horst Stürmer (eds.), *Politische Konzepte und verbale Strategien: Brisante Wörter – Begriffsfelder – Sprachbilder,* 149–167. Frankfurt: Peter Lang.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.

Vu, Hoang, George Keilas, Kimberly Metcalf & Ruth Hermann. 2000. The influence of global discourse on lexical ambiguity resolution. *Memory & Cognition* 28(2). 236–252.

Warnke, Ingo H. & Jürgen Spitzmüller. 2008. Methoden und Methodologie der Diskurslinguistik – Grundlagen und Verfahren einer Sprachwissenschaft jenseits textueller Grenzen. In Ingo H. Warnke & Jürgen Spitzmüller (eds.), *Methoden der Diskurslinguistik. Sprachwissenschaftliche Zugänge zur transtextuellen Ebene,* 3–54. Berlin: Mouton de Gruyter.

Warnke, Ingo H. & Jürgen Spitzmüller (eds.). 2008. *Methoden der Diskurslinguistik: Sprachwissenschaftliche Zugänge zur transtextuellen Ebene.* Berlin: Mouton de Gruyter.

Wengeler, Martin. 2003. *Topos und Diskurs: Begründung einer argumentationsanalytischen Methode und ihre Anwendung auf den Migrationsdiskurs (1960–1985)*. Tübingen: Niemeyer.

Wengeler, Martin & Alexander Ziem. 2014. Wie über Krisen geredet wird: Einige Ergebnisse eines diskursgeschichtlichen Forschungsprojektes. *Zeitschrift für Literatur und Linguistik* 173. 52–74.

Wengeler, Martin & Alexander Ziem (eds.). 2013. *Sprachliche Konstruktionen von Krisen: Interdisziplinäre Perspektiven auf ein fortwährend aktuelles Phänomen*. Bremen: Hempen.

Ziem, Alexander. 2009. Diskurse, konzeptuelle Metaphern, Visiotypen: Formen der Sprachkritik am Beispiel der Kapitalismusdebatte. *Aptum* 1. 18–37.

Ziem, Alexander. 2010a. Welche Rolle spielt der Kontext beim Sprachverstehen? Zum Stand der psycholinguistischen und kognitionswissenschaftlichen Forschung. In Peter Klotz, Paul R. Portmann & Georg Weidacher (eds.), *Kontexte und Texte: Soziokulturelle Konstellationen literalen Handelns*, 59–83. Tübingen: Narr.

Ziem, Alexander. 2013a. Wozu Kognitive Semantik? In Dietrich Busse & Wolfgang Teubert (eds.), *Linguistische Diskursanalyse: Neue Perspektiven*, 217–242. Wiesbaden: Springer.

Ziem, Alexander. 2013b. *Krise* im politischen Wahlkampf: linguistische Korpusanalysen mit AntConc. In Frank Liedtke (ed.), *Die da oben: Texte, Medien, Partizipation*, 69–90. Bremen: Hempen.

Ziem, Alexander. 2013c. Argumentstruktur-Konstruktionen als diskurslinguistische Analysekategorie. *Zeitschrift für Semiotik* 35(3–4). 447–470.

Ziem, Alexander. 2014. *Frames of understanding in text and discourse: Theoretical foundations and descriptive applications*. Amsterdam: Benjamins.

Ziem, Alexander. 2017. Wortschatz II: Quantifizierende Analyseverfahren. In Kersten Sven Roth, Martin Wengeler & Alexander Ziem (eds.), *Handbuch Sprache in Politik und Gesellschaft*. Berlin: Mouton de Gruyter.

Ziem, Alexander. In press. Korpuslinguistische Zugänge zur Lexik im politischen Sprachgebrauch. In Jörg Kilian, Thomas Niehr & Jörg Schiewe (eds.), *Handbuch Sprache und Politik*. Bremen: Hempen.

Ziem, Alexander, Ronny Scholz & David Römer. 2013. Korpusgestützte Zugänge zum öffentlichen Sprachgebrauch: Spezifisches Vokabular, semantische Konstruktionen und syntaktische Muster in Diskursen über "Krisen". In Ekkehard Felder (ed.), *Faktizitätsherstellung in Diskursen: Die Macht des Deklarativen*, 329–358. Berlin: Mouton de Gruyter.

Methodologies in sociolinguistic fieldwork

Adina Staicov
University of Zürich

Collecting data is a key aspect of linguistic research and a skill that requires more than simply going into the field and talking to people. This paper discusses the main stages of fieldwork and relates these to the author's own experiences as a fieldworker. From initial preparation and researching a community, through being in the field recruiting and recording participants, to returning home and managing the data, fieldwork needs to be planned carefully and meticulously in order to be a successful endeavour. Collecting spoken data is largely a social task and the right ways of talking to participants and conducting interviews can significantly influence the outcome of fieldwork. The aim of this paper is to show that, while collecting data is by no means an easy task, it is an important and valuable experience and a skill that sociolinguists wanting to work with their own data should acquire and refine.

1. Introduction

Fieldwork is the university of life – Nicole Eberle, p.c.
Fieldwork is an exercise in adaptability – Lena Zipp, p.c.
You gain so much more than just data by doing fieldwork – Marianne Hundt, p.c.

Fieldwork, as indicated by the above quotes by experienced fieldworkers, is a task that not only affects your professional, but also your personal life, as researchers do not only learn more about the community they are investigating, but also about themselves. Embarking on the journey to collect data for a project takes careful preparation, in-depth knowledge of the community one wants to study, and the willingness and adaptability to potentially move beyond one's comfort zone.

Based on my own experience conducting research I will point out some important issues that need to be kept in mind when planning, carrying out, and wrapping up the adventure that is fieldwork. I myself have been involved in field trips in different capacities: (i) as a student collecting data for the *International Corpus of English* (ICE); (ii) as part of a team conducting a pilot study; and (iii) collecting data for my own PhD project. The first fieldwork site was Fiji, an archipelago located in the South Pacific. The aim of this trip was to collect spontaneous conversations between Fijians and Indo-Fijians for the spoken component of ICE Fiji. Together with Lena Zipp, I conducted a pilot study with British Asians in London (Hundt & Staicov in preparation); the main focus of this study was to test the research design. And finally, my own PhD project focusses on identity construction in the San Francisco Chinatown community (Zipp & Staicov 2016; Staicov in preparation), which is where I spent a total of eight months to collect interview and discussion data. In what follows, I will draw on these experiences when elaborating on some of the key aspects of collecting data in the field.

2. The first steps: Preparing for fieldwork

2.1. Researching the community

The starting point of fieldwork begins long before travelling to the community that will be investigated. Together with consulting literature related to fieldwork (e.g. Bowern 2008; Sakel & Everett 2012; Schilling 2012) talking to experienced colleagues, supervisors or peers who have already conducted fieldwork is an excellent way of getting advice on where and how to begin.

Before designing a project, researchers first need to decide what they want to investigate and which community they want to study. Based on their interests or potential research gaps, researchers formulate questions that guide their study and that inform the choice of methodology. Research questions can be found by reading existing literature in one's field of interest, as authors often point towards avenues for future studies. As mentioned above, discussing topics with colleagues can also provide insight into interesting fields for research. Once a research question is formulated, the appropriate methodology to investigate the research question has to be found. Here, again, existing studies and communicating with other researchers can be useful. Finding the methodology that fits a research question is immensely important as it will influence the success or failure of data collection. A good way of testing a chosen method is a pilot study. The advantages of conducting such a, usually small scale, study will be discussed in Section 2.3.

Another important step is to get to know the community with which one wants to conduct research in as much detail as possible. Being familiar with the community's history and culture in general, but also with its demographics, with the geographic layout of the fieldwork site, with local events and issues, etc. will help the researcher when entering the community. As "all social events, including language use, are necessarily contextualized (spatially, temporally, historically, or otherwise) and potentially multivalent" (Levon 2013b: 69) a thorough understanding of the community is essential and will enable the researcher to draw informed conclusions later on. Existing research and more general literature on the community, the internet, census data, historical records, or travel guides will offer valuable insights into the field site and can also provide pointers as to how a community might best be accessed. All these resources should be consulted as early as possible in the planning process as it will potentially inform choices when deciding who to include in a sample, and who to better leave out.

During a seminar leading up to fieldwork in Fiji, for example, I and the other researchers involved focussed, among other issues, on the role of politics, religion, and gender. Religion plays a big part in the lives of many Fijians and it was therefore important to be respectful of participants' beliefs, even if we might not have shared them. Also, as gender roles are rather traditional, it made quite a difference to participants how especially the women in our group dressed and behaved and one female consultant commented very favourably on the fact that we were always dressed smart and neat. While it is likely that we would have been able to collect the data even without this knowledge, our sensitivity to these issues definitely resonated well with our participants.

2.2. How many participants to record?

After familiarising oneself with the community, deciding on a tentative number of participants is a good idea, as it will guide the researcher during the recruitment period. The number of participants depends on the research question and the number of social variables against which the linguistic variables will be tested. If the research question is based on a quantitative approach, for example, large numbers make a sample more reliable and representative. For qualitative research, an in-depth analysis of data of fewer participants can already yield valuable insights. A common standard, especially for quantitative studies, are five participants per social variable. Tagliamonte (2006: 31) herself, however, suggests two informants per social variable as the very minimum. When we conducted a pilot study in London, Lena Zipp and I decided to only collect data on eight participants. This number was sufficient for the purpose of our pilot study, which was to test our research design. For the actual study, however, the aim was to collect data from at least four times that number. As one aim of the project was to investigate both phonetic and morphosyntactic variation quantitatively, a larger number was necessary in order to allow for statistical testing, both with regard to participants and to extracted tokens of specific variables.

For my own PhD project in San Francisco Chinatown, I intended to collect interview and discussion data of at least 50 participants. After a total of eight months in the field, I only managed to collect 28 sets of recordings that I could use for analysis. The main reason for this lower number is that, while people were interested in the project, they were not willing to actually participate in the study. As researchers we must respect people's wishes and thus have to be flexible and adapt to the participants' and community's pace of life.

2.3. The pilot study

Testing one's research design before entering the field is helpful to detect potential shortcomings of the experiment and to test the equipment that will be used in the field (see Section 2.4). In the pilot study of which I was part, we tested the setup of our experiment, the set of questions prepared for both the discussion and the interview and, finally, our own interview skills. A pilot study can also shed some light on the time informants may have to invest when participating in a project, information that is clearly relevant to potential informants.

As a result of the pilot study we were able to make several adjustments to our study. In the pilot, we told participants that they could choose whichever topic they wanted during the discussion, but provided them with a list of everyday topics (music, current events, sports, etc.) that they could use as stepping stones. It turned out that this approach led to rather artificial conversations where people would go through the provided topics like a check-list, rather than have a more natural conversation (see also Rüdiger, this volume). For my PhD project I changed the topics to more specific questions that also focused on identity, one of the issues I was interested in. Despite the discussion being somewhat more restricted, fewer and more specific questions allowed my informants to have more meaningful conversations. We also noticed that the sound of watches or bracelets could interfere with the recordings, which can be particularly problematic for later phonetic analysis. Asking participants to remove such objects can be awkward, but is still better than ending up with data that cannot be used. Finally, we also learned more about how we performed as interviewers. It may seem straightforward to ask mainly simple, open-ended questions, but the pilot study taught us to focus more on how to better formulate such questions and also to be comfortable with a few seconds of silence.

In sum, it is evident that a pilot study can never fully prepare a researcher for all the difficulties that may arise in the field and, to a certain extent, failure is part of the experience. Nevertheless, a pilot study can help to gauge the suitability of the research design and allows the researcher to be as prepared as possible.

2.4. Equipment

It has become common practice to audio- or video-record informants when collecting spoken sociolinguistic data. Recordings allow the researcher to revisit the data, to store conversations in full, and to increase the accountability of one's choices with regard to data collection and analysis (Schreier 2013: 21). Furthermore, the "collection of sociolinguistic data in the form of recordings is absolutely crucial for quantification" (Schreier 2013: 21), which is becoming an integral part of the vast majority of linguistic research.

Depending on where fieldwork is conducted and how mobile the researcher needs to be, recording equipment has to be both high quality and practical. In all three settings where I conducted research, a small portable recording device, a H2 Zoom,[1] was used. As a rule of thumb, recording devices should be as inconspicuous as possible. Where practicable, a lapel microphone that can be attached to the researchers' or informants' clothes is a good option too, as it may further distract the informant from being recorded. With all the technological advances taking place at high speed, using a smart phone to record speech might be another valuable choice in the future. In order to ensure that no data is lost, it is paramount to familiarize oneself with the recording device and to carry not only an extra set of batteries, but also a back-up recorder. Furthermore, the device should be checked right before the recording to avoid empty or unintelligible files. Finally, all recordings should be saved on at least one more back-up disk as, speaking from personal experience, nothing is more frustrating than having to realize that a recording has been lost.

Now that the researcher is familiar with the community they want to study, that a number of participants has been agreed upon, and that the appropriate recording device has been chosen, it is time to go to one's research site and to enter the community.

3. In the field: Collecting data

3.1. Entering the community

Contacting potential participants in the field is the single most important aspect in conducting fieldwork. No matter how prepared a researcher is, if there are no informants, there is no project. The researcher's status vis-à-vis the community, that is, as an insider or outsider, calls for different strategies when recruiting participants. For a researcher who is an insider, contacts are already existent and data collection can start almost immediately. As Hoffman (2014: 31) states, "[y]ou are a true participant observer, with a natural place in the community". Furthermore, as the

[1] The H2 Zoom is a handheld device that can be used to record mono and stereo, and two- or four-channel sound wav or mp3 audio files. With its relatively small dimensions it is easy to carry and handle and can be placed rather inconspicuously. A newer type, H2n Zoom with improved features is now available.

researcher has intimate and first-hand knowledge of the community, finding appropriate topics or issues to talk about in an interview should be less problematic (see Section 3.2). However, as an insider, issues of objectivity might arise or, as Levon (2013a: 202) points out, "[t]he difficulty in this type of research [...] is making the transition from 'regular community member' to 'researcher' as smooth as possible". This means that insiders need to make an effort to move beyond their own social network in order to capture a broader, more representative sample of the community under investigation.

For researchers who are outsiders, establishing initial contacts might be more difficult and different strategies can be adopted when trying to find these first contacts. A well-known method in sociolinguistics is the friend-of-a-friend approach first described by Milroy (1980). Here, a researcher is introduced to the community via a middleman, or -woman, a contact who functions as door opener and who introduces the researcher to their social network. Similar to the problems described for the insider above, the friend-of-a-friend method might limit the scope and diversity of the sample and it might be necessary for the researcher to try and access members who are located more on the periphery of the target network. Nevertheless, this approach is immensely valuable as the researcher is integrated into the community through somebody with whom participants are familiar and who they trust. Furthermore, depending on the field site, staying with locals can be a great opportunity to meet people and to get acquainted with the community at large (Schreier 2013: 25). However, it seems that this approach is better suited for smaller, more rural communities. In a large city like San Francisco, where anonymity can be a challenge, this approach is certainly not impossible, but more difficult.

If the researcher does not have the option to use the friend-of-a-friend approach, contacting professional stranger handlers (Agar 1996) is another valid option. These professionals can be teachers, community leaders, religious authorities, government officials, etc.; they are familiar with the community, and therefore can help establish exchanges between the fieldworker and members of the community. Using brokers to enter a community comes with a caveat that should not be underestimated, namely, that the broker might choose informants based on how they represent the community. This pre-selection could lead to data that is not representative of the wider speech community. The fieldworker needs to be aware of this limitation and should, again, try to move beyond the initial contacts in order to record a broader variety of participants.

For the different projects in which I was involved, we used a mix of the above-described methods. In Fiji, where no previous contacts had been established, we organized a get-together early on in our stay to introduce ourselves and the project to the university community that represented our target community. On the one hand, we were able to make appointments with students who agreed to participate, and many of these first participants then referred us to their friends. On the other hand, we met with members of staff who allowed us to attend their lectures and seminars in order to further promote the study. This combined approach of using brokers and the friend-of-a-friend method proved very successful and allowed us to collect more than 100 spontaneous conversations during a three-week stay. In London, participants were recruited by a PhD student at University College London where we were based. Using the university mailing list, our local contact selected suitable candidates for our pilot study, which allowed us to complete the study in a relatively short amount of time. Finally, in San Francisco, I contacted Chinatown-based organisations, local colleges, universities and schools, posted flyers in the neighbourhood, and attended different community events. As I did not have any previously established contacts in the community, recruiting informants proved immensely challenging and I had to try many different avenues before I could record my first participant. One relatively successful method was the distribution of flyers, and it was also through this method that the first participant was recruited. The majority of informants, however, were found by attending service at church. While only a few members of the congregation originally participated, I was able to eventually tap into their friendship circle, and the above-mentioned friend-of-a-friend method proved to be most effective for this community. This experience resonates with the introductory quote on adaptability, as I had to reconsider my methods on the spot in order to finally arrive at my goal of actually recording people.

Despite the setbacks in the beginning, trying different methods proved successful in the end and, in combination with the experience described for Fiji and London, should encourage future fieldworkers to remain positive and persistent, and to exhaust different possibilities of entering a community.

Finally, it is important to point out that in the three projects mentioned above, the data sets were collected based on judgment sampling, meaning that informants were chosen based on predefined categories rather than randomly. The latter approach is useful for large-scale projects, where participants are selected based on phone books or census data. For the projects described here, this approach would not have been feasible or appropriate as the studies followed very specific criteria or research questions. Depending on the research questions, different sampling methods might be necessary and should also be decided on in advance.

3.2. Recording conversations

The type of data one needs to collect depends, again, on the specific project and research questions one has in mind. In general, fieldworkers aim to collect natural free speech that closely resembles the vernacular, the mode of speaking people use when they are most relaxed and pay least attention to their speech (Labov 1972). In order to capture different levels of formality (from casual to highly formal), the sociolinguistic interview (Labov 1984) is an approach employed by many sociolinguists (see also Rüdiger, this volume). A traditional sociolinguistic interview consists of: (a) a casual interview that focusses on the informants personal experiences; (b) a guided interview that allows for increased comparability across samples; (c) a reading task; (d) a word list; and (d') a list of minimal pairs (see e.g. Labov 1984 for a detailed description of the sociolinguistic interview). However, it is not necessary to always collect all the parts that make up the sociolinguistic interview and often emphasis is placed on the more casual part of the interview (see Becker 2013). In Fiji, for example, the aim was to collect informal discussions between two informants; in London, we collected interview data, discussions, and speech produced in a goal-oriented task; in San Francisco, the focus was again on interview and discussion data.

While recording a conversation might sound relatively straightforward, attention needs to be paid to the set-up of the interview, to the location where the conversation is being recorded, and to the effect that the researcher might have on the informants' speech. Finally, research ethics have to be adhered to at all times. In order to keep track of the data collection process and to note down observations on the community or on linguistic features, keeping field notes is a practical way of recording and storing all the information in one place.

The interview set-up

Recording a person presents an intrusion into their life as many informants share their personal opinions and experiences with the researcher. In order to be as considerate of the informants as possible, the researcher should be aware of potential pitfalls and challenges that might occur during the interview. Similar to everyday conversations, certain topics (such as politics, migration, ethnicity) might not be suitable for the research context. Here, the researcher's knowledge of the community (see Section 2.1) should be exploited as, based on this information, the researcher should have an idea of what to talk about, and what to avoid. If a taboo topic is raised accidentally, the researcher should apologize and steer the conversations into another direction.

Interviews might be free in that no specific topic must be touched upon, but even in this situation, having a list of topics at hand can be useful in case the conversation comes to a halt. For studies where the content of an interview is not an integral part of analysis, any topic will serve the purpose of data collection and the conversation can flow naturally. If the researcher is interested in studying a particular issue in the community, for example in ethnographic fieldwork, pre-defined questions need to be asked in order to be able to collect relevant data. In such a situation, it is important to find a balance between applying pre-defined concepts (etic) of an issue to the conversation, and letting the participant provide local interpretations (emic) of a particular matter. A combination of an etic with an emic approach seems the most favourable, as

> too much reliance on practitioners' own understanding can push ethnography toward an untenably strong relativist position, one that gives individuals free range as agents and fails to recognize the larger social, institutional, and ideological forces that shape interaction. (Levon 2013b: 70)

For my PhD project in San Francisco, I had prepared a set of questions that I wanted to discuss with my informants ranging from question on their (family's) migration history to life in Chinatown, or language practices; a selection of questions is provided in (1)–(4).

(1) When did your family come to the US (San Francisco)?
(2) What does it mean to you that San Francisco has a Chinese American mayor?
(3) Have you ever got comments on the way you speak?
(4) What role does language play for your identity?

By asking specific questions I provided informants with my thoughts on certain topics, for instance asking the question in (4) can mean that I believe language to play a role for a person's identity. At the same time, however, my questions gave participants the opportunity to critically discuss these topics and to add to or correct my assumptions. This approach allowed me to steer the conversation in a certain direction but also to give a lot of control to the participants. Furthermore, it added to my understanding of the community and of issues that were relevant to my participants and allowed me to adjust my approach to better fit my research. Adaptability during the data collection period is vital and "it is [...] important to remain open to the possibility that things are not how you originally imagined them to be when you were still an outsider" (Levon 2013b: 71).

While it is preferable to have a conversation with a consultant rather than a question-answer situation, the researcher should remember that the focus is on the consultant. This does not mean that a fieldworker cannot contribute to the conversations, but their input should be limited and interruptions or talking over participants should be avoided. A researcher should be interested in the participants' stories and not be afraid to let this show. Using common sense and courtesy can go a long way and authenticity is better than pretending to be someone one is not. From my own experience I find that informants are happy to talk about many topics as long as they feel that the fieldworker's interest in the topic and the participant is genuine.

One of the biggest challenges when recording speech features under the label "Observer's Paradox". As Labov describes:

> the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation. (Labov 1972: 209)

It is very unlikely to fully distract a participant from the presence of a recording device. Nevertheless, it is the task of the fieldworker to try to make informants feel at ease during the interview setting so as to minimize the effect of the Observer's Paradox as much as possible. Researchers can, for example, meet with the informants in a location of the participants' choice if this does not negatively affect the quality of the recordings (see below). Talking about topics the participants are interested in or sharing some information about oneself can also help break the ice and make the interview setting less artificial.

The interview itself is the central part of data collection and researchers should spend some time to learn as much about how to conduct an interview as they can, and they should practice conducting interviews before going into the field. As mentioned above, participants might share some very personal information with the researcher, which puts them in a more vulnerable position. Being aware of one's own limits (for example as regards one's safety or potential issues that might complicate data collection, see below) and being considerate of the participants' situations are thus key skills of a field worker or, as Schreier describes, "good fieldworkers know when and where to get data; excellent fieldworkers know when and where not to get data" (Schreier 2013: 27).

The location

The quality of the recording is immensely important for later analysis, especially if the focus is on phonetic features. Where possible, a quiet space indoors is the best option to record, as ambient noise is reduced.

On Fiji, many recordings were made outdoors and a lot of outside noise like birds singing, other people talking, or traffic was caught on the audio. This kind of prominent background noise made transcribing files more difficult and, consequently, complicated data analysis. In London, we had the luxury of recording in sound-proof rooms, which was an advantage in terms of audio quality. However, a disadvantage of this location was that the setting was very artificial and created a formal atmosphere, aspects that might have affected the participants' speech (see Rüdiger, this volume). In San Francisco, I was able to collect the majority of my data in a small office space that I had rented, and these recordings were the most easy to work with at the analysis stage. However, this setting again created an atmosphere that was somewhat more formal.

Deciding on a suitable location can result in a trade-off between comfort for the participant and quality of the audio material. In such a case, I would argue that the participant's comfort is more important as everybody is more likely to talk when they feel comfortable rather than inhibited by an artificial, laboratory-like environment.

The role of the interviewer

Researchers themselves are another factor that could influence the outcome of an interview situation. As discussed in Section 3.1, the position of the fieldworker vis-à-vis the community can already establish certain power asymmetries, especially if the researcher is an outsider. However, as others have pointed out "the real authority always lies with the person who *provides* the data, not the one depending on them" (Schreier 2013: 28, emphasis original). This view echoes Schilling-Estes (2007: 181), who stated that, "highlighting one's role as a learner and the role of participants as experts on their community can get a long way toward obtaining casual speech and building good relations."

The researcher's appearance can also be an influencing factor and in order to "minimi[ze] dominance" (Schilling-Estes 2007: 181) dressing more casually is another way of making informants feel more comfortable around the fieldworker. I have already provided an example on how clothing can have a positive effect in Section 2.1. I have made similar experiences both in London and in San Francisco. Working with students in London was more relaxed when I was wearing more casual clothes, stressing that the age difference between us was small. In San Francisco, I dressed more formally when I attended church service or for more official meetings, but later adapted to the participants who were very casual. It is also important to feel comfortable as a researcher as this will already make a somewhat artificial situation more relaxed.

While dress is a relatively easy factor to control, personal characteristics like age, gender, or ethnicity are not (see e.g. Cukor-Avila & Bailey 2001 for a discussion on the role of race). The majority of encounters I had with informants were very pleasant and none of the features mentioned above proved to be problematic. However, it was clear in certain situations that gender roles, for example, were defined differently to what I was used to, a circumstance that affected some of the recordings in Fiji. In this rather traditional society, some male informants used the discussions to record their phone numbers or to ask some fieldworkers out on a date. While such examples were rare, fieldworkers need to be aware of potential difficulties that might arise as a result of such cultural differences and, in potentially difficult cases, might consider the possibility of having members of the community collect the recordings. Societies that are highly racialized could lead to challenges as well, as the race or ethnicity of the fieldworker might be assessed negatively by informants. As Schilling-Estes (2007: 182) states:

> Despite society's best efforts to eliminate prejudices and discrimination based on factors such as sex, ethnicity, and age, people will form certain opinions about researchers and perhaps limit their access to community groups and interactions based on such factors.

Luckily, I was never in a situation where I felt that my persona was problematic. However, this is not to say that it did not affect the recordings at all. Similar to the Observer's Paradox, it is unlikely that researchers will be able to distract from the fact that they are conducting research, no matter how relaxed the interview setting is. While this does not mean that the collected data cannot be used, the role the researcher plays should not be underestimated and needs to be addressed when discussing potential limitations of the research design.

Research ethics

The final and one of the most important points I want to mention are research ethics and the researcher's responsibility towards the participants. Participating in a research project means sharing personal information with an individual or institution with which one is not very familiar. While many people do not object against sharing their stories with a researcher, many feel more comfortable knowing that their identity is protected and that all potential clues to their identity will be anonymized. This is especially true in vulnerable communities and "we must be mindful so that [participants] are secure in their anonymity and our promise of confidentiality" (Hoffman 2014: 29).

Researchers should also be as transparent as possible about what they want to study and how. With regard to language, however, too much information could have a negative effect on participants and in linguistic research it is thus important to find a balance between too little and too much information. Too much detail will draw the informants' attention to the linguistic feature analyzed, which means that participants might try to avoid producing relevant features because of self-consciousness, for example. Thus, it is important to prepare a consent form where participants are informed about the general intent of the study. With an informed consent form, researchers can show that they follow good practice and that they have not collected their data surreptitiously. By signing such a form, informants agree to participating in the study and to their data being used for linguistic research. Even with such consent forms, it is unlikely that participants will be fully aware of what a linguistic study entails and how their data will be analyzed (see Appelbaum et al. 1982; Miller & Bell 2002 for more detailed discussion of this issue). Nevertheless, using a consent form shows that the researcher is aware of the ethical implications of working with personal data and that they have asked participants for permission to use the collected data. Additionally, fieldworkers should always provide contact information so that informants can reach them to obtain more information after data collection is completed, or to withdraw consent if they so wish. Finally, researchers might also want to think about how to give back to the community. This might not always be easy but small gestures will be appreciated by all participants. In the London and San Francisco settings, informants were paid for participating in the respective studies. Some object to this practice as it can, for example, put pressure on researchers who often struggle for funding, or it can create expectations in communities so that people only participate if they are remunerated (see e.g. Schreier 2013). Despite these concerns I realized that it was the only incentive that would allow us to attract enough participants. Having said that, I believe that many people will be happy to participate in a study and that money should thus be the last resort. Offering help, or a copy of the final product of the study, or some small token from your country are other ways of giving back and are equally appreciated (see also Rüdiger, this volume). Ways of giving back should be addressed in the planning process and more experienced researchers are a valuable source for advice on this matter.

4. Back home: Sorting and storing data

Once data collection is done and the researcher has returned home, the data needs to be sorted and stored properly. Considerable work can be saved by organising the data during the collection period, especially when several recording sessions are scheduled in a row and time for back-ups and labelling might be tight.

In a first step, researchers should decide on a simple way of labelling their data including information on the speakers, the date the recording was made and, potentially, some important information on the interview setting. At this stage, the identity of the participant should already be protected. Pseudonyms are a good way of anonymising information, alternatively numbers or letter codes can be used. For the latter option, combinations should follow a specific pattern in order to be as simple as possible. For our London participants we included information on a participant's heritage language and a number that represented the order of the recording: H02 thus means heritage language = Hindi, second person recorded out of the Hindi cohort = 02. If informants are recorded in different settings or if different types of data are collected, the label identifying a consultant should be the same across all data sets. Field notes will help make sense of the data and add relevant information on the recordings, for instance short descriptions on specific files.

Ideally, the researcher has already made back-ups of the data during fieldwork. If this is not the case, creating back-up files is one of the first things that should be done upon return. Even if such a file already exists, it is a good idea to create a second one on a different device, as nothing is more frustrating than losing carefully collected data. Storing large amounts of audio and video data has become easier and also more affordable with new technology and having digital copies also allows for a quick search of data.

Depending on the type of data and the research question, the data might have to be transcribed. While it is beyond the scope of this paper to discuss transcription procedures and conventions (see e.g. Jenkins 2011; D'Arcy 2013), researchers should again include information on the informants, interviewer, recording, etc. in the transcript and use the same label again to identify the participants.

Keeping track of one's data can be tedious but is paramount to both the data collection process and to the sorting and storing process afterwards. Field notes can be used as a mnemonic device if labelling and sorting have been neglected but the better the system is designed from the beginning, the easier it will be to work with the data back home.

5. Conclusion

Fieldwork is both an exciting and demanding endeavour that takes careful preparation, enthusiasm for meeting new people, and resilience to deal with potentially challenging situations. It was the aim of this paper to illustrate the different stages of fieldwork – preparation, data collection, and post-processing – and to provide some insights into how these different stages can be approached. Fieldwork largely depends on the people involved, on the characters of the fieldworker as well as the participants. In some cases the two will not be compatible and it will be better to move on to the next participant, but in many cases, the encounters a fieldworker makes will be valuable professionally and, maybe even more so, personally. Keeping an open and curious mind will help fieldworkers through more difficult periods of data collection, which will turn into cherished anecdotes once fieldwork has been successfully completed. Schreier (2013: 18) argues "against conventional wisdom, that fieldwork is not that easy after all" and most linguists who have conducted fieldwork will agree. Fieldwork is not easy and it cannot be done on the spot. However, with the right preparation, research on the community, with adaptability and persistence it will be an enriching experience a researcher will gain much more from than "just" a good set of data.

References

Agar, Michael H. 1996. *The professional stranger: An informal introduction to ethnography*. New York: Academic Press.

Appelbaum, Paul S., Loren H. Roth & Charles Lidz. 1982. The therapeutic misconception: Informed consent in psychiatric research. *International Journal of Law and Psychiatry* 5(3-4). 319–329.

Becker, Kara. 2013. The sociolinguistic interview. In Christine Mallinson, Becky Childs & Gerard van Herk (eds.), *Data collection in sociolinguistics,* 91–100. New York: Routledge.

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. Basingstoke: Palgrave Macmillan.

Cukor-Avila, Patricia & Guy Bailey. 2001. The effects of the race of the interviewer on sociolinguistic fieldwork. *Journal of Sociolinguistics* 5(2). 254–270.

D'Arcy, Alexandra. 2013. Vignette 11c. Advances in sociolinguistic transcription methods. In Christine Mallinson, Becky Childs & Gerard van Herk (eds.), *Data collection in sociolinguistics,* 187–190. New York: Routledge.

Hoffman, Michol. 2014. Sociolinguistic interviews. In Janet Holmes & Kirk Hazen (eds.), *Research methods in sociolinguistics: A practical guide,* 23–41. Malden: Blackwell.

Hundt, Marianne & Adina Staicov. In preparation. Expressing and negotiating identity in the London Indian Diaspora: Towards the quantification of qualitative data.

Jenkins, Christopher J. 2011. *Transcribing talk and interaction: Issues in the representation of communication data*. Amsterdam: Benjamins.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh & Joel Sherzer (eds.), *Language in use*, 28–83. Englewood Cliffs: Prentice Hall.

Levon, Erez. 2013a. Ethnography and recording interaction. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 195–215. Cambridge: Cambridge University Press.

Levon, Erez. 2013b. Ethnographic fieldwork. In Christine Mallinson, Becky Childs & Gerard van Herk (eds.), *Data collection in sociolinguistics*, 69–80. New York: Routledge.

Miller, Tina & Linda Bell. 2002. Consenting to what? Issues of access, gatekeeping and 'informed' consent. In Melanie Mauthner, Maxine Birch, Julie Jessop & Tina Miller (eds.), *Ethics in qualitative research*, 53–69. London: Sage.

Milroy, Lesley. 1980. *Language and social networks*. Baltimore: University Park Press.

Sakel, Jeanette & Daniel L. Everett. 2012. *Linguistic fieldwork: A student guide*. Cambridge: Cambridge University Press.

Schilling, Natalie. 2012. *Sociolinguistic fieldwork*. Cambridge: Cambridge University Press.

Schilling-Estes, Natalie. 2007. Sociolinguistic fieldwork. In Robert Bayley & Ceil Lucas (eds.), *Sociolinguistic variation: Theories, methods, applications*, 165–189. Cambridge: Cambridge University Press.

Schreier, Daniel. 2013. Collecting ethnographic and sociolinguistic data. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 17–35. Cambridge: Cambridge University Press.

Staicov, Adina. In preparation. "I feel like I'm a born again Chinese – but kinda fake": Ethnic identity construction in San Francisco Chinatown's diaspora community. Zurich: University of Zurich dissertation.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Zipp, Lena & Adina Staicov. 2016. English in San Francisco Chinatown: Indexing identity with speech rhythm? In Elena Seoane & Cristina Suárez-Gómez (eds.). *World Englishes: New theoretical and methodological considerations*, 205–228. Amsterdam: Benjamins.

Cuppa coffee? Challenges and opportunities of compiling a conversational English corpus in an Expanding Circle setting

Sofia Rüdiger
University of Bayreuth

Geographic variation of the English language provides a plethora of research opportunities for linguists. Long gone are the days when the focus of those studies was solely on native English speaking countries; more recent research does not only take second language varieties into account but inquires also into English spoken as a foreign language (EFL). Most investigations of structural features in EFL contexts rely on written material, whereas studies using spoken material are rarer. Spoken data is connected to many challenges when it comes to data collection, processing and analysis, but nevertheless offers insights into basic processes of language change. This paper introduces the "cuppa coffee" data collection method employed to collect a corpus of spoken English by South Korean speakers. I will show how the simple act of framing the sociolinguistic interview as new acquaintances drinking a cup of coffee together helps to avoid a language learning and teaching framework, puts participants in a more relaxed mindset and finally results in more "naturalistic" and richer conversational data. The framing relies heavily on social conventions of coffee drinking and capitalizes on the status of coffee as a "social lubricant".

1. Introduction

Methodological reflection on the (ethnographic) interview process has been likened by Briggs (1983: 233) to "opening the Pandora's box". It is not my intention to open a Pandora's box. However, in this paper I want to re-engage in a reflection process regarding a specific data collection method in linguistics: the sociolinguistic interview. Methodological issues are often brushed aside, since "[c]ollecting data is viewed as an intrinsically sound, if not necessarily glamorous pursuit" (Briggs 1986: xiii). Nevertheless, as Briggs (1986: xiv) noted "it is worth setting aside preconceptions regarding the triviality of 'purely methodological' questions". Methodological reflection deserves considerable attention in works on the variation of English. Nowadays, the inclusion of a methodology section can be considered standard in empirical publications on language. The description of the actual data collection process, however, often only receives minimal attention, further shown by its length: One or two pages are common for a full-length monograph. This seems little, especially considering the actual amount of time and effort involved: Fieldwork often takes up several weeks or months (or even years) and is usually preceded by a lengthy process of finding and deciding on the data collection method, designing research instruments and in many cases performing pilot studies (see also Staicov, this volume).

Many linguists have been concerned with the attainment of vernacular, casual or informal speech vis-à-vis the performance of speech in the interview situation. Labov (1966: 99) emphasized the necessity to elicit this "everyday speech which the informant will use as soon as the door is closed behind us: the style in which he argues with his wife, scolds his children, or passes the time of day with his friends". This notion is based on the thought that interview speech is constrained by the perceived formality of the situation and it is the linguists' desire to find a way around these constraints (commonly featuring under the label "Observer's Paradox"). Labov himself advocated several means for this, such as letting the interview participants[1] tell "brink of death" stories, group interviews or surreptitious observation of the participants (used in his famous department store study). How adequate these methods are in attaining very intimate speech styles has been criticized for example by Labov himself (1972: 90) and Rickford (1987: 153). Rickford (1987: 154), however, concedes that "the Observer's Paradox is real, its major supporting principles sound, and the need for techniques like spontaneous interviews and group recordings indisputable".

My goal is not to dissuade linguists from using interviews as a data source for their research. On the contrary, my own research on spoken English by Korean speakers heavily draws on a self-compiled corpus of spoken interactions between myself and the research participants. However, I encourage variational linguists, especially those using spoken

[1] I do not distinguish between the terms participants, informants and interviewees in this paper and will use them interchangeably.

material, to put more emphasis on methodological issues and devote more space in their writing on reporting methodology.

As I will explicate in Section 4.3, especially in EFL as well as English as a Second Language (ESL) contexts, attention must be paid to interview framing. In this paper, I will introduce the "cuppa coffee" frame which, even though simple in conception, I found extremely helpful in the data collection process (depending of course on the research questions and the overarching goal of the study). I want to start out by introducing the status of spoken and written data in linguistic research in general and in World Englishes studies more specifically (Section 2). Methodological discussions often hinge upon the differentiation between natural and contrived data, and I will provide a succinct summary of the debate (Section 3). In the following, I will introduce the cuppa coffee method and how it works (Section 4). Finally (Section 5), I will present the results of my own cuppa coffee research activity – the corpus of **Spo**ken **K**orean **E**nglish (SPOKE) – followed by some concluding remarks (Section 6).

2. Spoken and written data in linguistic research

We can identify a bias towards written language in English linguistic research in general, and Miller (2006: 671) even claims that "most published work on English deals with the written language". For the field of applied linguistics (but also linguistics in general), McCarthy (1998: 16) reports the use of written material as "baseline data" at the expense of spoken language research. The same bias becomes evident in variational linguistics when taking a closer look at one of the most prominent journals in the field: *World Englishes*. Surveying the articles published in the journal in the years 2014 and 2015 (excluding two special issues), reveals that only 8 out of 42 articles (19%) are based either partly or completely on spoken language data.[2] Breaking these numbers down according to which types of varieties were investigated in the respective article, the situation becomes even more dire for Expanding Circle Englishes: 6 out of 20 papers (30%) on Inner and/or Outer Circle varieties are based at

[2] Excluding spoken interviews eliciting material on language attitudes, etc.

least partly on spoken language data (largely due to the availability of material from the *International Corpus of English* [ICE]), but only 2 out of 14 Expanding Circle English studies (14%) use spoken data as material for analysis.[3] And in both of these cases, the language investigated is mediated language, more specifically the language of television, which – even though spoken in nature – is, depending on the type of format, often at least semi-scripted and can usually not be taken as a basis for studies on conversational language use.

Collecting spoken data can be riddled with methodological (and ethical) pitfalls: Telling participants that they are being recorded sparks the notion of the Observer's Paradox (i.e. that people behave differently when they know that they are being observed; see Section 1), but not telling the subjects that they are being recorded violates basic guidelines of ethical research.

Many researchers aim to balance these issues by telling their participants that they are being recorded, but not what speech forms the researcher is interested in. Thus, the participants are left uninformed of the actual research interest, which potentially could pertain to all linguistic levels (pronunciation, word choice, discourse strategies, morphosyntax, etc.). Even though this arguably evades ethical issues, it does not completely override the Observer's Paradox, as speakers might still be more likely to monitor their speech. Additionally, the collection and processing of spoken data takes more time and can be described as tedious at best and in the long run, "wickedly expensive" (McCarthy 1998: 12).

Apart from methodological and ethical issues, researchers also have to deal with analytical difficulties in their work with spoken material. Spoken language is neither simple nor disorganized; its complexity lies within its "dynamic and intricate" character (Halliday 1990: 87; see also Biber et al. 1999: ch. 14). More disfluencies in spoken language (such as incomplete sentences, repetitions, hesitation markers, pauses) do not only make automatized procedures more difficult, but also complicate manual data analysis.

[3] The complete breakdown of the numbers according to variety status is as follows: 20 Inner/Outer Circle varieties (6 of those based on spoken language), 14 Expanding Circle varieties (2 of those based on spoken language), 3 Inner/Outer/Expanding Circle varieties (none of those based on spoken language), 5 unclear.

However, the collection and analysis of spoken data not only involves many challenges, but also offers numerous opportunities. Of particular importance for the field of variational linguistics is the fact that “[m]any constructions begin life confined to spoken language but make their way into writing” (Miller 2006: 679). In other words, spoken language can be deemed the “motor of language change” (Kortmann 2006: 615). For studies on varieties, especially emerging varieties, which investigate very dynamic linguistic situations, spoken data is thus a valuable resource to “catch” language change on the go or emerging patterns. This has also been acknowledged by other scholars in the field, such as Schneider (2004: 247), who ratifies that “oral performance is less constrained and less conservative than written styles, so this is where innovations are most likely to surface”. But not all spoken data is created equal and the distinction between so-called natural and artificial data has been at the heart of many methodological discussions.

3. Natural vs. artificial data

The difference between natural (also referred to as “naturally occurring” or “naturalistic”, see Speer 2002a: 513) and artificial (also referred to as “non-naturally occurring”, “researcher-provoked” or “contrived”, see Speer 2002a: 513) spoken data has received considerable scholarly attention. This dichotomy regarding spoken data is applied in methodological descriptions of studies despite the fact that “the status of pieces of data as natural or not depends largely on what the researcher intends to ‘do’ with them” (Speer 2002a: 513). As Speer (2002a: 514–515) continues to elucidate, the research traditions of conversation analysis and discursive psychology have a strong preference for naturally occurring data. Even scientists in the aforementioned disciplines accede that all interactions are prone to bias and context-dependency.

The presumption that “naturally occurring talk is ‘better’ than contrived materials, or more amenable to analysis, because it would have happened ‘had the researcher not been born’” (Speer 2002a: 516) continues to prevail. Nevertheless, it remains unclear how researchers are supposed to obtain natural data if natural data is speech produced as if the researcher did not exist and the conversation is therefore not being investigated (and thus also recorded). As Speer (2002a: 521) points out, “natural data” is a keyword in linguistic research methodology which “has become a catch-all term with fuzzy boundaries and little in the way of specificity”. Speer describes, for example, how she used picture prompts in her research on gender views during otherwise naturally occurring dinner conversations with her friends and family (Speer 2002b: 547). Despite the dinner conversations occurring without being artificially triggered, the use of the picture prompts manipulates the conversation to take a specific direction (i.e. the discussion of gender), which leads Speer to wonder whether this data can be clearly categorized as either “natural” or “artificial”.

Potter (2002: 541) suggests distinguishing between “natural” and “naturalistic” data by using a “(conceptual) dead social scientist test – would the data be the same, or be there at all, if the researcher got run over on the way to work?”. An interview can, of course, not happen without the interviewer being there (and alive), whereas, for example, “a counselling session would take place whether the researcher turns up to collect the recording or not” (Potter 2002: 541). Naturally occurring talk thus needs to be “produced entirely independently of the actions of the researcher” (Potter 2011: 190).

Comparing narratives told in interviews and spontaneous conversation, Koven (2011: 87) found that “interview stories may be as interactionally complex and amenable to interactional analysis as conversational stories”. In other words, the results from the interview situation were very similar to the results in the conversational setting, which challenges the notion of the interview as an artificial speech event whose analytic results are going to deviate from the “real thing”, that is, naturally occurring speech. However, Koven (2011: 88) adds (in a footnote) that “[t]here are different types of conversation, as there are different types of interviews”. Taking a closer look at the interview setting of the reported study shows that even though the interviewer and the interviewee did not know each other before the interview, they were of a similar age and shared the same autobiographic background (i.e. being “raised in France by Portuguese migrants”; Koven 2011: 76). These factors might have helped to create a less intrusive and more relaxed atmosphere. In fact, it is this kind of atmosphere that linguists need to establish in order to elicit more naturalistic data. In the following I want to depict how this can be achieved with a very simple approach: the cuppa coffee method.

4. The cuppa coffee method

Keeping my research questions in mind,[4] I formulated the criteria for the data I needed at the outset of my research on spoken English in South Korea (henceforth, Korea): The material should be conversational in nature, authentic and as natural as possible. As my research focused exclusively on morphosyntactic patterns (e.g. the use of plural morphemes, articles, tenses, pronouns, etc.), sound quality was not of utmost importance but still needed to be good enough to discern whether participants used, for example, word final *-s* as a plural morpheme. The data should also stem from a non-classroom context and needed to be collected within the realms of ethical research. Regarding authenticity, I discarded the idea of having Korean speakers talk to each other without me being present, but instructing them to use English beforehand. Since the unmarked language choice for Koreans communicating with other Koreans is always Korean, putting the participants into such a situation would have created a highly artificial speech event. Talking to me, a stranger from Germany, however, predisposes the use of English as this is the unmarked language choice when talking to foreigners.

One of the communicative situations where Koreans would actually use English when talking to each other is in specific classroom contexts. I, however, especially wanted to avoid evoking this language learning and teaching mindset. Education in general and English language education more specifically holds a momentous position in Korean society (see e.g. Seth 2002 on "education fever" in Korea and J. K. Park 2009 on "English fever" in the Korean society; for attitudes towards English see J. S.-Y. Park 2009). English proficiency tests (e.g. TOEFL [Test of English as a Foreign Language], TOEIC [Test of English for International Com-

[4] The overarching research questions for the project described were: Does the English spoken by Koreans show morphosyntactic variation? If yes, is this variation systematic?

munication], NEAT [National English Ability Test] and TEPS [Test of English Proficiency]) are extremely high-stakes issues for many Koreans[5] and are thus connected with anxiety, insecurity and uneasiness. Due to those rather negative connotations and for the sake of a more authentic conversation with my participants, I wished to avoid a classroom and language learning context.

After some deliberation, I decided to not only invite my informants to an interview in exchange for a cup of coffee but to actively frame our encounters as new acquaintances meeting for coffee in a café. The basis for my cuppa coffee method is thus the [CUP OF COFFEE] frame. It inherently marks sharing a cup of coffee as a social activity that necessarily includes conversation.

4.1. Why coffee?

Coffee (and tea) are ideal candidates for this conversational framing as they have been identified as "social beverage" (Hattox 1985) and "social lubricant" (Grund 1993; Valeri 1996). The act of drinking a cup of coffee together creates "a favorable atmosphere for communication through the mutual participation in the coffee or tea drinking ritual" (Grund 1993) and drinking coffee is seen as a "stimulating, comforting, [...] sociable and talkative" (Valeri 1996: 139) activity. These notions are "exploited" by the cuppa coffee method which invites the interviewees to participate in a social activity coined by its social character.[6] Additionally, as already indicated by the reference to the "coffee or tea drinking ritual"

[5] English is, for example, part of the university entrance exams as well as a requirement on recruiting exams by companies (see Lee 2006: 67). English thus functions as a status symbol in Korean society (Shim & Baik 2004), is the "key to upward social mobility" (J. S.-Y. Park 2009: 37), and learning English has even been compared to a national religion (J. S.-Y. Park 2009: 1).

[6] It needs to be kept in mind, though, that drinking coffee has two sides: It can be an inherently social activity, as it is portrayed in this paper, but it can also be a solitary activity, for example, when drinking coffee at one's desk at work. We can even find this aspect of solitude in cafés (seemingly buzzing with activity and conversation), where individual coffee drinkers can enjoy a beverage by themselves in "calculated copresence" (Varnelis & Friedberg 2008: 17) with the other patrons and without verbally engaging with each other. The default expectation when going to a café in company with someone else, though, is for social interaction to occur.

in the previous quote, coffee drinking is a ritualized process (Anderson 2003: 165) and thus coupled with specific expectations regarding participant behaviour. A supporting role for the cuppa coffee method is played by the notion of the coffee manipulation: Social psychologists have found that simply holding a cup of coffee (or any other warm/hot beverage) generates "feelings of interpersonal warmth" (Williams & Bargh 2008: 606) such as trust and comfort.

As Gaudio (2003: 660) explains, an invitation for a cup of coffee indexes a specific speech situation, that is, "a scheduled, informal, face-to-face encounter between ostensible social equals in a coffeehouse or other commercial catering establishment". From an American point of view the conversations which ensue in these contexts can be characterized as "'casual', 'ordinary' or even 'natural'" (Gaudio 2003: 660). Cafés are actively constructed in the media and by café owners as "space[s] of interaction" (Gaudio 2003: 675), whose arrangement of architecture, furniture and lighting is conducive to conversation in general (Gaudio 2003: 682). Although drinking coffee did not have this connotation from the outset of its career, the phrase "'Let's have a cup of coffee' came to mean 'Let's have a conversation'" (Topik 2009: 99).

Asian countries[7] are traditionally connected to tea-drinking rather than coffee-drinking, and especially Japan is renowned for its complex cultural, highly ritualized traditions of tea-serving and tea-drinking (see e.g. Cross 2009; Surak 2013). Up to the 1960s, coffee was mainly consumed in North America and Europe, but after that coffee consumption started to boom in Asia as well, particularly in Japan and in Korea (Daviron & Ponte 2008[2005]: 150). Coffee was first introduced to Japan by the Dutch in the 1690s with the first coffee shops opening in 1888 (Ueshima 2013: 197). Cook and Lee (2008: 88) show how coffee-drinking in China is "already showing every sign of a strongly entrenched habitus presence", despite a strong tea-drinking tradition. Coffee carries a connotation of Western ideology and is strongly connected to modernization and "urban sophistication" (Cook & Lee 2008: 84). Importantly, it is characterized as "gregarious", relationship- and trust-building (Cook & Lee

[7] For a geographical overview of coffee vs. tea consumption see Grigg (2002).

2008: 91). In Korea, "just over 100 years since the introduction of coffee to Korean society, coffee drinking has become an important part of Korean food culture" (Bak 2005: 38). The USDA Foreign Agricultural Service (Coffee Market Brief Update 2015) reports steady growth of coffee industry resources in South Korea and specifies that the coffee consumption per head is five times higher than in the other countries in the Asia-Pacific region. According to the report, coffee is more popular than tea and can be seen as an "established beverage" on the Korean market. As one Korean newspaper, The Korea Herald, avidly summarized it: On average, Koreans consume coffee more often than they eat kimchi (a traditional side dish of fermented, spicy cabbage) or rice (Chung 2015). Coffee drinking in Korea is generally associated with global modernity and Koreans have been described as "regular drinkers" of coffee (Bak 2005: 39). In the following, I want to describe the [CUP OF COFFEE] frame and how it can be utilized as a tool in the interviewing process.

4.2. Framing

In her discussion of data collection methods for sociolinguists, Schilling distinguishes between factors that can be controlled by the researcher, such as the questions asked or the setting of the interview, and aspects which cannot be controlled by the researcher, "such as how participants choose to frame the research situation, [...] e.g. as a casual conversation vs. a formal informational [interview]" (Schilling 2013: 128). Indeed, the way participants frame the interview cannot be controlled to the same degree as, for example, the interview setting, but I argue that it is within the interviewer's possibilities to "guide" the framing of the interview into a more informal direction by framing the interview process him/herself in a certain way. Of course, it has to be kept in mind that whether this framing is beneficial to the study underway depends largely on the research goals in general (i.e. does the researcher strive to attain conversational data in the first place).

Frames have been discussed by cognitive psychologists (e.g. Tversky & Kahneman 1981), sociologists (most prominently Goffman 1974) and linguists alike and defining the term itself is already a challenging endeavour (Keren 2011). Other terms in use are *scene* (e.g. Fillmore 1975), *schema* (e.g. Tannen & Wallat 1993), *script* (e.g. Schank & Abelson 1977) and *scenario* (e.g. Sanford & Garrod 1981), which, as Bednarek (2005: 688) explains, differ in terms of emphasis rather than conceptually. I take the term *frame* to refer to "a process whereby communicators, consciously or unconsciously, act to construct a point of view that encourages the facts of a given situation to be interpreted by others in a particular manner" (Kuypers 2009: 182). In linguistics, the term frame has frequently been used to describe how language is used to achieve or support a certain frame, the effects of framing on narratives of past experiences (e.g. Kamoen et al. 2015; Dayter & Rüdiger 2016) and to explain the effects of expectations on linguistic output (Tannen 1978). In the case of the cuppa coffee method, however, I take the term frame out of the linguistic context and shift it more to the contextualization of the interview situation (even though this is of course also expressed in my verbalizations regarding the situation, e.g. I tried to refer to the situation as "a conversation", "a chat", etc. instead of "an interview"; nevertheless, I occasionally did use the term interview). Table 1 summarizes the roles, actions, objectives and props associated with the [INTERVIEW] and [CUP OF COFFEE] frames.[8] "Mixing" both frames results in the cuppa coffee method, which capitalizes on the direction of shift of roles, actions and objectives from [INTERVIEW] to [CUP OF COFFEE] frame, while employing props from both frames (i.e. recording device and beverage).

As we can see when comparing the [INTERVIEW] and the [CUP OF COFFEE] frame both activities are inherently connected to speech.[9] But in the case of interviews, we find a rather hierarchical set of fixed rules as to who might say what, whereas the conversation during a cup of coffee seems to be more egalitarian in nature. One of the most essential "rules" of the interview speech event in fact is the question-answer format, which is so finely ingrained that an interviewer's attempts at holding a free conversation "will usually arouse surprise and may even lead to suspicion and resentment" (Wolfson 1976: 190). Another rule of the

[8] The [CUP OF COFFEE] frame described in Table 1 refers to the social activity of drinking a cup of coffee and not its application as a data collection method.

[9] The two sides of drinking coffee, either social or individualistic in nature, have been previously described in footnote 6.

interview speech event signifies that only the interviewer, not the interviewee can introduce a new topic (Wolfson 1976: 192) and that it is the interviewer's call when the shift from one topic to the next is performed (see also Briggs 1984: 21). The natural roles held by the interlocutors outside the interview are therefore backgrounded and instead replaced by the roles of interviewer and interviewee, respectively (Briggs 1986: 2). Briggs makes this "interview frame" responsible for the participants' awareness "that messages will be decoded (and should be decodable)" (1984: 24).

Table 1. Comparison [INTERVIEW] and [CUP OF COFFEE] frame

	[INTERVIEW]	[CUP OF COFFEE]
Roles	Interviewer and interviewee	Conversational partners; speaker A and speaker B
Actions	Interviewer asks questions; interviewee answers questions	Both speakers drink beverage; both speakers participate in conversation equally (depending on individual character and conversational style)
Objective	Supply the interviewer with answers/data	Partake in social activity
Props	Recorder, microphone, list of questions	Beverage

Briggs found out during his research on wood carving in a New Mexican speech community that the interviews he prepared were rather fruitless whereas the (presumably tape-recorded) "conversations which were structured by the [observed] couple turned out to be extremely fertile" (1984: 23). A similar thing happened when the interviewer (Briggs) and his participants were talking while communally engaged in wood carving, which shows that it can be highly worthwhile to break out of the traditional interview framework.

Shifting from a structured interview to what is often called a "spontaneous interview", that is, an interview which is of a more conversational nature, has been described as "an exceedingly uncomfortable thing to do" (Wolfson 1976: 195). Whereas structured interviews provide both the interviewer and the interviewee with clear speech event rules and result in language which is natural for an interview situation, it has been claimed that the foraging into spontaneous interviews deviates from the expected speech event to such a degree that it can confound interviewees, resulting in artificial and unnatural language patterns. Wolfson (1976: 196) recounts an anecdote where a researcher, who was trying to set up a conversational interview, was "sent home" and told to return with better prepared questions by the participant. The conclusion here was that

> If speech is felt to be appropriate to a situation and the goal, then it is natural in that context. The context itself may be formal or informal, interview or conversation. It is only when norms of speaking are uncertain or violated that one gets 'unnatural' speech data. (Wolfson 1976: 202)

This clearly demonstrates the need to frame the interview as a conversational format from the *very outset* of the interaction. Framing the interaction as new acquaintances having a coffee together overrides the issues mentioned by Wolfson (1976), who claims that the missing "rules" for a spontaneous/casual interview speech event can confuse, irritate or even anger the interviewee. Having a talk over a cup of coffee is a speech event familiar to all (or at least most) adult speakers of industrialized and modern societies. It therefore overrides the missing rules regarding the casual interview and puts the speakers in a speech situation which is comfortable by its very own nature and its familiarity to the speakers.

The activation of the [CUP OF COFFEE] frame depends on three factors: the setting, the topics and the interviewer's behaviour. The obvious choice for the setting is of course a café. For the interviews conducted in my study, it was the participants who decided in which part of Seoul and in which café we should meet. During the first contact (usually via text message or e-mail) I suggested that participants pick a café that they liked and which was convenient for them to reach. No list of questions or topics was prepared beforehand and the encounter usually consisted of a “getting to know each other”. I usually let the conversation take its course and tried to identify either areas of common interest or topics which the participant was passionate about (whether it was philosophy or the difference between Korean and German sausages). I included some “sneaky” questioning though and tried to inquire at least once into future plans (e.g. plans for the next vacation or the coming weekend) and past experiences (e.g. travel abroad). My own speech behaviour can be described as talkative and when participants asked me questions (which occurred very frequently, see Section 5), I answered them in detail.

4.3. Advantages and drawbacks of the cuppa coffee method

Apart from getting data that is conversational in nature and more natural than staged sociolinguistic interviews, the cuppa coffee method helps to put participants in more of a story-telling mood, arguably due to the relaxed atmosphere of the conversation. The whole research setup was also very conducive to the “snowball”-method of participant acquisition, as my interlocutors were usually very keen to introduce me to their friends, family members or co-workers. Finally, the cup of coffee that I bought represented at least a small token of appreciation for the time and effort of my participants. I furthermore want to emphasize the non-deceptiveness of my research method: Yes, I framed the interviews explicitly in a light which was conducive to my research goals, but all speakers knew that they were being recorded and I attained informed signed consent from them before commencing the audio recording.

Actively constructing the encounter between researcher and participant as new acquaintances enjoying a cup of coffee together also works against the activation of a teaching and learning frame. Meeting participants, for example, in seminar rooms or university offices, always establishes an interviewer-interviewee hierarchy, and it will be very difficult to get participants out of an “I’m being tested”-mindset. The avoidance of the teaching and learning frame is especially important in EFL contexts due to speakers’ intense (prescriptively coined) contact with English in public and private schooling.[10] Using a language during a test situation is inevitably different from using the same language in a conversation and it is thus important to prevent falling into this mode (even unwittingly). Hadikin (2014: 41), for example, portrays his collection of spoken English by Korean speakers as follows: The participants of the study were "sitting at a table or desk, alone with myself [Hadikin] as the interviewer in either a classroom, small interview room or seminar room". Before recording, participants had approximately 15 minutes time to fill in a questionnaire with their demographic information as well as to answer more general questions, such as "How long have you been studying English? [...] What do you do in your spare time? Please write a short note about any of the following: [f]avourite TV/films [,] [a]ny kind of reading you like [,] [f]avourite music" (Hadikin 2014: 41). Hadikin (2014: 41) concedes that the setting might have been conceived by participants as "more formal [...] than was intended" and that the use of the written questionnaire resulted in language forms deviating from "purely spoken interaction", but in the end he characterizes his data still as "a close representation of 'natural' conversation". However, the whole setup of the interview situation portrayed closely resembles one of a language proficiency test and it is thus questionable how conversational the spoken material is after all.[11]

The large technical advances in recording technologies over the last decades of course also contribute to a successful implementation of the cuppa coffee method. Whereas it was difficult to place bulky tape recorders and microphones unobtrusively in front of interview participants at the outset of sociolinguistic studies, the recording devices nowadays are small, at times only marginally larger than mobile phones. This mitigates one dilemma of earlier times, where researchers frequently had to balance issues of data quality with issues of the unobtrusiveness of data collection (as described by McCarthy 1998: 12 and Tannen 2005: 43).

[10] Which is of course not to say that inquiring into this speech style is not worthwhile.

[11] This does not necessarily devalue Hadikin's results on collocational patterns in English spoken by Koreans.

Obvious drawbacks of the cuppa coffee method are the money spent on coffee (even though remuneration of participants via other means may of course be equally or more expensive). It should also be mentioned that the cuppa coffee method takes considerably longer than traditional interviews: The overall talking time on average was three times as long as the actual recording time. In hindsight, the interviews can be divided into three stages. Stage 1 consisted of ordering the coffee, finding a table and getting settled into our seats. As most Korean cafés are of the self-serve kind, we usually had to wait a couple of minutes until our beverages were prepared and then had to pick them up from the counter. During this waiting time, I already engaged in conversation with my interviewees. After picking up the beverages, I usually turned on the audio recorder (which can be labelled as stage 2). After approximately half an hour I turned off the audio recorder again but without breaking off the conversation. I usually continued to talk to my participants (stage 3) in the same way as we did during the "audio recorder is on"-stage for some time (on average around 30 minutes) until I or the interviewee had to go.

On the recordings I also find lengthy stretches of my own (i.e. interviewer) speech, which obviously cannot be used for the analysis of Korean English, but nevertheless had to be transcribed in order to be able to identify potential priming effects (and to paint a coherent picture of the conversations). Indeed, around 45% of the material collected (as determined by a word count) consists of my own speech, which shows the nearly equal distribution of turns between me and my interlocutors (at least regarding the amount of words uttered). I thus interpret this more as a measure of the success of my method than a "real" drawback (even though this factor has to be kept in mind when calculating the time needed for transcribing and the target number of words). A further (potential) drawback of the data collection procedure described here relates to the quality of recordings: Cafés are simply noisier locales (due to other café patrons talking, music playing) than quiet office spaces, seminar rooms or the phonetician's recording booth. Therefore, the data I collected via the cuppa coffee method may not be adequate, for example, for detailed acoustic analysis, but was viable for my research purposes (see Section 4).

Last but not least, the effects of high coffee consumption by the researcher should not be forgotten, who has to drink a cup of coffee during each interview (decaffeinated coffee might be a solution here for many regions of the world).

5. The Spoken Korean English Corpus (SPOKE)

My own application of the cuppa coffee method resulted in a collection of 60 hours of audio material by 115 speakers, which after transcription translates to roughly 300,000 words (excluding my own contributions to the conversations, which were transcribed for analytical purposes, but of course do not figure into the final word count of the corpus; see Section 4.3). Even though I tried to invite all of my participants to have a coffee with me in a local coffee shop, this was not always possible. If participants wished to meet at another locale, I followed their directions and in a few cases interviewed participants in their homes, in their office, a seminar room at university or even outside in a park. 26 of the 115 speakers were thus recorded in non-café settings, even though I still made sure in those cases that we both had a hot beverage in our hands. Still, 230,000 words (roughly 77% of the material) were collected following the cuppa coffee method. The coffee shops as such varied and I only infrequently visited cafés more than once with different participants. Some conversations were recorded in stores of larger coffee shop chains (e.g. *Dunkin Donuts, Krispy Kreme, Angelinus*; interestingly, only two recordings were made at *Starbucks*), whereas others took place in more local, privately owned cafés around Seoul.

One of the advantages of using the data collection method described can be found, for example, in the abundance of questions directed by the Korean participants to their interlocutor (i.e. the interviewer). The typical role assignment in interviews regulates the use of questions as follows: The interviewer asks the questions whereas the interviewee answers them. The interviewer may ask follow-up questions until s/he deems the question sufficiently answered and then moves on to the next question or topic. Questions by the interviewee are not budgeted for in this research format. I see the frequent occurrence of questions from the participants' side as a sign that the cuppa coffee method worked: A conversation consists of all participants contributing in more or less equal ways.[12]

Due to the presence of the tape recorder and the researcher, the interaction might not be "natural" as such, but we can claim at least that it is conversational to a certain degree (even though, in the end, there is no definite way to tell how relaxed the participants were during the recordings). Framing the interview as a conversation allows participants to ask questions themselves, which they frequently did in my data: They enquired into many aspects of my personal and professional life, and I tried to be very forthcoming in answering their questions. As some kind of by-catch of my methodological approach, this allows me now to look at the issue of question construction of Korean speakers of English. A more traditional sociolinguistic interview would have yielded far fewer questions by the interviewees and thus would have made this line of linguistic inquiry difficult if not even impossible.

6. Conclusion

Thirty years ago, Briggs (1986: 16) criticized sociolinguists' tendency to treat matters of methodology in general, regarding interviews specifically, "in passing" and not to devote more thought on methodological issues. Now, more than ever, it is adequate to re-think methodological approaches to spoken data collection since interviews commonly used by researchers to elicit informal or casual speech are problematic both in ENL (English as a native language) and EFL contexts. Whereas in the former this is the case due to the perceived formality of the situation by the informants, it is especially in the latter contexts that interviews can quickly "slip" into a language proficiency test situation, which is of course also detrimental to the attainment of casual conversational data.

In this paper I have shown how the framing of interviews as sharing a cup of coffee can be beneficial for linguistic research. I do not suggest that everybody should grab a cup of coffee with their research partici-

[12] Of course, this is only possible if the researcher is also willing to open up about his/her own life. Expecting interviewees to open up about their life, the interviewer needs to lead by example. Otherwise it will be even more difficult to establish interviewer and interviewees as equals.

pants; however, the cuppa coffee framework can be conducive for many interview situations. Constraints of culture and social context have to be considered carefully though (a slew of questions along the following lines come to mind here: Is it appropriate in this culture to have a cup of coffee with a stranger? How about a stranger from the opposite sex? How old are the research participants? Is it still appropriate to invite them for a cup of coffee?). Depending on the context, it might be more appropriate to find another social framework connected to verbal interaction into which the interview could be transplanted.

We currently still lack studies comparing the data collected in artificial interview situations with material from natural conversations (see Koven 2011 for one of the rare examples), which will further our reflection process on methodological issues related to spoken data immensely. Studies documenting the understanding of the different roles in interviews in different cultures will also be helpful in this regard. For an instrument heavily employed in sociolinguistic research, surprisingly little is still known about the interview as a speech event per se.

No data collection method is inherently perfect and the appropriateness of a research instrument always depends on the nature of the research and the related research question(s). Methodological issues are hard to come to terms with, but we should not lose track of the means available to us as linguists: Our knowledge of speech events and human communication in general should help us when it comes to the sociolinguistic interview as a research method. And sometimes, something as simple as drinking a cup of coffee with the participants can make a difference.

References

Anderson, Eugene N. 2003. Caffeine and culture. In William Jankowiak & Daniel Bradburd (eds.), *Drugs, labor, and colonial expansion*, 159–176. Tucson: The University of Arizona Press.

Bak, Sangmee. 2005. From strange bitter concoction to romantic necessity: The social history of coffee drinking in Korea. *Korea Journal* 45(2). 37–59.

Bednarek, Monika. 2005. Frames revisited: The coherence-inducing function of frames. *Journal of Pragmatics* 37. 685–705.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Briggs, Charles L. 1983. Questions for the ethnographer: A critical examination of the role of the interview in fieldwork. *Semiotica* 46(2/4). 233–261.

Briggs, Charles L. 1984. Learning how to ask: Native metacommunicative competence and the incompetence of fieldworkers. *Language in Society* 13(1). 1–28.

Briggs, Charles L. 1986. *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge: Cambridge University Press.

Chung, Joo-Won. 2015. Koreans consume more coffee than kimchi, rice. *The Korea Herald*. http://www.koreaherald.com/view.php?ud=20150116000639 (6 May, 2016.)

Coffee Market Brief Update – USDA Foreign Agricultural Service. 2015. http://gain.fas.usda.gov/Recent%20GAIN%20Publications/Coffee%20Market%20Brief%20Update_Seoul%20ATO_Korea%20-%20Republic%20of_12-31-2015.pdf (5 May, 2016.)

Cook, Jackie & Robert Lee. 2008. The espresso revolution: Introducing coffee-bar franchising to modern China. In Lawrence C. Rubin (ed.), *Food for thought: Essays on eating and culture*, 83–96. Jefferson: McFarland & Company.

Cross, Tim. 2009. *The ideologies of Japanese tea: Subjectivity, transience and national identity*. Folkestone: Global Oriental.

Daviron, Benoit & Stefano Ponte. 2008. What's in a cup? Coffee from bean to brew. In David Inglis, Debra Gimlin & Chris Thorpe (eds.), *Food: Critical concepts in the social sciences*, 130–169. London: Routledge.

Dayter, Daria & Sofia Rüdiger. 2016. Reporting from the field: The narrative reconstruction of experience in Pick-up Artist online communities. *Open Linguistics* 2(1). 337–351.

Fillmore, Charles. 1975. An alternative to checklist theories of meaning. In *Proceedings of the first annual meeting of the Berkeley Linguistics Society, Institute of Human Learning*, 123–131. Berkeley: University of California.

Gaudio, Rudolf P. 2003. Coffeetalk: Starbucks^{TM} and the commercialization of casual conversation. *Language in Society* 32(5). 659–691.

Goffman, Erving. 1974. *Frame analysis: An essay on the organization of experience*. New York: Harper & Row.

Grigg, David. 2002. The worlds of tea and coffee: Patterns of consumption. *GeoJournal* 57(4). 283–294.

Grund, Jean-Paul. 1993. The concept of ritualization. http://www.drugtext.org/Drug-Use-as-a-Social-Ritual/2-the-concept-of-ritualization.html (11 November, 2015.)

Hadikin, Glenn. 2014. *Korean English: A corpus-driven study of a new English*. Amsterdam: Benjamins.

Halliday, Michael A. K. 1990. *Spoken and written language*. Oxford: Oxford University Press.

Hattox, Ralph. 1985. *Coffee and coffeehouses: The origins of a social beverage in the medieval Near East*. Washington: University of Washington Press.

Kamoen, Naomi, Maria B. J. Mos & Willem F. S. Dekker (Robbin). 2015. A hotel that is not bad isn't good: The effects of valence framing and expectation in online reviews on text, reviewer and product appreciation. *Journal of Pragmatics* 75. 28–43.

Keren, Gideon. 2011. On the definition and possible underpinnings of framing effects: A brief review and a critical evaluation. In Gideon Keren (ed.), *Perspectives on framing*, 3–33. New York: Psychology Press.

Kortmann, Bernd. 2006. Syntactic variation in English: A global perspective. In Bas Aarts & April McMahon (eds.), *The handbook of English linguistics*, 603–624. Malden: Blackwell.

Koven, Michele. 2011. Comparing stories told in sociolinguistic interviews and spontaneous conversation. *Language in Society* 40(1). 75–89.

Kuypers, Jim. 2009. Framing analysis. In Jim Kuypers (ed.), *Rhetorical criticism: Perspectives in action*, 181–204. Plymouth: Lexington Press.

Labov, William. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lee, Jamie Shinhee. 2006. Linguistic constructions of modernity: English mixing in Korean television commercials. *Language in Society* 35(1). 59–91.

McCarthy, Michael. 1998. *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Miller, Jim. 2006. Spoken and written English. In Bas Aarts & April McMahon (eds.), *The handbook of English linguistics*, 670–691. Malden: Blackwell.

Park, Jin-Kyu. 2009. 'English fever' in South Korea: Its history and symptoms. *English Today* 25(1). 50–57.

Park, Joseph S.-Y. 2009. *The local construction of a global language: Ideologies of English in South Korea*. Berlin: Mouton de Gruyter.

Potter, Jonathan. 2002. Two kinds of natural. *Discourse Studies* 4(4). 539–542.

Potter, Jonathan. 2011. Discursive psychology and the study of naturally occurring talk. In David Silverman (ed.), *Qualitative research: Issues of theory, method and practice*, 187–207. Los Angeles: Sage.

Rickford, John R. 1987. The haves and have nots: Sociolinguistic surveys and the assessment of speaker competence. *Language in Society* 16(2). 149–178.

Sanford, Anthony & Simon Garrod. 1981. *Understanding written language*. Chichester: Wiley.

Schank, Roger C. & Robert Abelson. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge*. Hillsdale: Lawrence Erlbaum.

Schilling, Natalie. 2013. *Sociolinguistic fieldwork*. Cambridge: Cambridge University Press.

Schneider, Edgar. 2004. How to trace structural nativization: Particle verbs in world Englishes. *World Englishes* 23(2). 227–249.

Seth, Michael J. 2002. *Education fever: Society, politics, and the pursuit of schooling in South Korea*. Honolulu: University of Hawai'i Press.

Shim, Rosa Jinyoung & Martin J. Baik. 2004. Korea (South and North). In Ho Wah Kam & Ruth Y. L. Wong (eds.), *Language policies and language education: The impact in East Asian countries in the next decade*, 172–193. Singapore: Eastern Universities Press.

Speer, Susan A. 2002a. 'Natural' and 'contrived' data: A sustainable distinction? *Discourse Studies* 4(4). 511–525.

Speer, Susan A. 2002b. Transcending the 'natural'/'contrived' distinction: A rejoinder to ten Have, Lynch and Potter. *Discourse Studies* 4(4). 543–548.

Surak, Kristin. 2013. *Making tea, making Japan: Cultural nationalism in practice*. Stanford: Stanford University Press.

Tannen, Deborah. 1978. The effect of expectations on conversation. *Discourse Processes* 1(2). 203–209.

Tannen, Deborah. 2005. *Conversational style: Analyzing talk among friends*. Oxford: Oxford University Press.

Tannen, Deborah & Cynthia Wallat. 1993. Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview. In Deborah Tannen (ed.), *Framing in discourse*, 57–76. Oxford: Oxford University Press.

Topik, Steven. 2009. Coffee as a social drug. *Cultural Critique* 71. 81–106.

Tversky, Amos & Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211(4481). 453–458.

Ueshima, Tatsushi. 2013. Japan. In Robert W. Thurston, Jonathan Morris & Shawn Steiman (eds.), *Coffee: A comprehensive guide to the bean, the beverage, and the industry*, 197–200. Lanham: Rowman & Littlefield.

Valeri, Renée. 1996. Coffee in Sweden: A social lubricant. In Jonas Frykman & Orvar Löfgren (eds.), *Force of habit: Exploring everyday culture*, 139–150. Lund: Lund University Press.

Varnelis, Kazys & Anne Friedberg. 2008. Place: The networking of public space. In Kazys Varnelis (ed.), *Networked publics*, 15–42. Cambridge: MIT Press.

Williams, Lawrence E. & John A. Bargh. 2008. Experiencing physical warmth promotes interpersonal warmth. *Science* 322(5901). 606–607.

Wolfson, Nessa. 1976. Speech events and natural speech: Some implications for sociolinguistic methodology. *Language in Society* 5(2). 189–209.

Detecting and discouraging non-cooperative behavior in online experiments using an acceptability judgment task[1]

Jana Häussler and Tom Juzek
University of Wuppertal and University of Oxford

Crowdsourcing is an attractive means for data collection. It is cheap, fast, and has a broad demographic coverage, but it is also susceptible to non-cooperative behavior (i.e. participants are not complying with the task). Our study examines the impact of non-cooperative behavior in linguistic studies using an acceptability task and recruiting participants via *Amazon's Mechanical Turk*. Data from twelve experiments show that non-cooperative behavior does not result in zero-mean noise but affects the results in a substantial way. In our data, we identified three types of non-cooperative participants: simple spammers (who "click their way through", without giving meaningful ratings), clever spammers (they also click their way through, but make a few pauses that make their overall times look normal), and inattentive participants (normal response times but poor ratings). While simple spammers stand out by showing extremely short response times, a second type of spammer is harder to catch. Yet, a median-based response time criterion detects them as well. Inattentive participants can be identified by their performance on specific items ("booby trap items"). Further, we demonstrate that implementing a warning mechanism that tracks response times and produces a warning when they repeatedly fall below a certain threshold reduces non-cooperative behavior considerably.

1. Introduction

Within theoretical linguistics, and in particular within syntax, there is an ongoing debate on the empirical foundation of linguistic theories (among others Schütze 1996; Newmeyer 2003; Wasow & Arnold 2005; Phillips 2010; Gibson et al. 2011; Sprouse et al. 2013). Grammaticality judgments are a main data source for syntactic theories, but they are often collected in a way that is rather informal, that is, it does not adhere to common standards in related disciplines, such as cognitive psycho-

[1] We thank two anonymous reviewers for their insightful comments as well as the audience of the MaLT conference for a lively discussion.

logy. There are often just a few lexicalizations for each condition, all versions of an item are presented together, and so on. The number of participants is very low and includes the researcher. Such “armchair linguistics” (which conflates the role of informant and researcher) has been criticized as potentially biased and unreliable. Consequently, over the past twenty years, the number of studies using formal empirical methods has considerably increased. And with the development of crowdsourcing platforms such as *Amazon*'s *Mechanical Turk*, it has become even easier to collect data in an easy, cheap, and fast way. Further, crowdsourcing allows researchers to sample from a broader population than most lab studies (though still not fully representative of speech communities). It therefore does not come as a surprise that more and more linguists are using crowdsourcing techniques for their studies (e.g. Schnoebelen & Kuperman 2010; Gibson et al. 2011; Sprouse 2011b; Sprouse et al. 2013).

1.1. Data quality

Despite the obvious benefits of crowdsourcing, worries about data quality persist. Dandurand et al. (2008) raise concerns about the number of uncontrolled variables (noise, distraction, technical equipment, etc.), multiple submissions by the same person, a higher dropout rate, and self-selection biases. Arguably, however, the most serious concern is reliability. Can we trust the data? Do participants comply with the task and work as carefully as in the lab?

Previous studies showed that crowdsourcing can be as reliable as lab experiments (e.g. Krantz & Dalal 2000; Dandurand et al. 2008; for linguistic studies see Munro et al. 2010; Schnoebelen & Kuperman 2010). Other studies, though, have demonstrated that crowdsourcing is quite susceptible to non-cooperative behavior, that is, participants not complying with the task (e.g. Downs et al. 2010; Kazai et al. 2011). Kazai et al. (2011), for instance, estimate that up to 57% of their participants were non-cooperative (depending on how “non-cooperative” is defined exactly; see Section 1.2).

Our study adds to the discussion by (i) testing the reliability of linguistic acceptability judgment tasks in which participants were crowdsourced and (ii) examining the effectiveness of three strategies to detect and prevent non-cooperative behavior.

1.2. Non-cooperative behavior

We consider a participant's performance non-cooperative when he/she does not comply with the task. Note that we use the term "non-cooperative" in a broad sense, including unintentional non-compliance due to factors like distraction and fatigue. (The study presented below involves linguistic acceptability judgment tasks, so the non-cooperative behavior that we observe mainly consists of submitting weak ratings at fast response times; in extreme cases, we observe quasi-random ratings at unrealistically fast response times.)

Non-cooperative behavior would be a minor nuisance if it just created zero-mean noise, that is, a random variable with an expected mean of zero.[2] In practice, zero-mean noise would increase variance but not affect the mean. Ratings would randomly deviate from the mean but cancel out each other if averaged. To test whether non-cooperative behavior is indeed that harmless, we compare experimental results including versus excluding problematic participants (for tools to detect non-cooperative behavior see Section 3). Our results show that non-cooperative behavior is anything but harmless: It affects the data in a serious way by creating non-zero-mean noise.

Hence, we need to detect and exclude participants with non-cooperative behavior. Relevant strategies are discussed in Section 3. However, it would be even better if we could prevent such behavior. Section 4 discusses criteria which can be used to bar potentially non-cooperative participants from taking part in an experiment in the first place. Section 5 introduces an effective technique for discouraging non-cooperative behavior during the experiment.

[2] An example of zero mean noise is Gaussian white noise. Each sample of such noise has a normal distribution with a zero mean.

2. A breakdown of our study

Our study comprises three experiments in which we crowdsourced our participants (the experiments were not designed to examine non-cooperative behavior in the first place, but to pursue other methodological questions; for details see Juzek 2016 and Häussler & Juzek 2015).[3] Participants were recruited using *Amazon*'s *Mechanical Turk*. The actual experiments were run on an external website. Each experiment was run in four sessions, making a total of twelve sessions. For each session, we recruited 40 participants, adding up to a total of 480 participants. Payment corresponded to an hourly rate of about $10. Only native speakers of American English were included in the analyses, which led to an exclusion of 18 participants in Experiment 1, 23 participants in Experiment 2, and 22 participants in Experiment 3. This leaves us with 417 participants for analysis (including non-cooperative participants).

Table 1. An overview of our experimental sessions

Experiment	Method	Modality	Items	Participants
1 a	5pt Likert scale	Auditory	64	40
		Visual	64	40
b	5pt Likert scale	Auditory	64	40
		Visual	64	40
2 a	7pt Likert scale	Visual (part 1)	108	40
		Visual (part 2)	108	40
b	Binary	Visual (part 1)	108	40
		Visual (part 2)	108	40
3 a	Binary	Visual	40	40
b	7pt Likert scale	Visual	40	40
c	Thermometer judgment	Visual	40	40
d	Magnitude estimation	Visual	40	40

[3] See http://tsjuzek.com/resources/Haeussler_Juzek_CLS_Extended_Lab_Or_Armchair.pdf for an extended abstract.

The experiments employed four types of scales: binary judgments, judgments on a 5-point or 7-point scale, thermometer judgments (Featherston 2008) and magnitude estimation (Sorace 1992; Bard at al. 1996; Cowart 1997). In each case, the participants' task was to judge the acceptability of sentences that appeared on the screen, except for Experiment 1, in which sentences were presented auditorily in two of the four sessions. For an overview of the procedures see Table 1.

2.1. Experiment 1

Experiment 1 investigated differences between spoken and written English. The materials included two constructions that mainly occur in spoken language (resumptive pronouns and alternative *if*-clauses) and two constructions that mainly occur in written language (sentence-initial gerunds and *wh*-infinitives). Each construction was represented by four lexicalizations. Examples are given in (1) to (4).

(1) We are afraid of things that we don't know what they are.
(2) If she would come to see things for herself, she would change her mind immediately.
(3) Their being unaware of the situation really annoyed Rob.
(4) We found a splendid house in which to spend our holiday.

In addition, Experiment 1 included eight sentences each from a spoken source (*National Public Radio*) and a written source (mainly from *USA Today*). Half of the sentences in each set were modified to decrease their acceptability (change in agreement, drop of function, etc.).[4] Finally, 32 fillers were added including ungrammatical ones. Four items were used for a calibration phase at the beginning of the questionnaire (for details see Section 3.2). In total, Experiment 1 examined 64 sentences. The sentences were identical in all four sessions.

Experiment 1 had two sub-experiments to control for a possible confounding factor, viz. formality (spoken language is typically associated with an informal register whereas written language typically coincides with more formal situations; for details see Juzek 2016: ch. 2).

[4] The materials are available oline at www.tsjuzek.com/thesis_additional_materials.

2.2. Experiment 2

Experiment 2 compared acceptability judgments obtained in an informal versus formal way. To this end, Experiment 2 collected acceptability ratings for 200 sentences randomly sampled from a corpus of sentences that occurred in articles published in *Linguistic Inquiry* (LI) between 2001 and 2010. The study is related to Sprouse et al. (2013), but we sampled single sentences instead of sentence pairs. In total, the study included 100 sentences marked with an asterisk in the original articles and 100 unmarked, that is, acceptable, items (for details, see Häussler & Juzek 2015). Experiment 2a used a 7-point scale (ranging from 1, "fully unacceptable", to 7, "fully acceptable"). Experiment 2b used binary judgments (unacceptable/acceptable). Within each sub-experiment, the 200 sentences were distributed across two sessions so that each participant rated 100 sentences from the corpus (50 marked sentences and 50 unmarked sentences). The order of items was randomized for each participant individually.

(5) John turns out to be winning. (Becker 2006: 451)
(6) *John believes Mary to hit Bill. (Martin 2001: 163)

In addition, Experiment 2 included the four calibration items used in Experiment 1. Furthermore, four "booby trap" items were included to filter out inattentive participants (for details see Section 3.2).

2.3. Experiment 3

Experiment 3 examined data transformations for four different types of scales for collecting acceptability judgments: (i) yes-no judgments (Experiment 3a), (ii) a 7-point scale (3b), (iii) a self-anchoring scale known as thermometer judgments (3c; cf. Featherston 2008), and (iv) a magnitude estimation task (3d; cf. Stevens 1946, 1951).[5]

With the method of thermometer judgments, participants define their own reference points – a minimum (denoting "fully unnatural/ungrammatical") and a maximum ("fully natural/grammatical"). They then rate

[5] That is, in contrast to Experiment 3a and 3b, Experiments 3c and 3d had no preset scale.

sentences with respect to these points. The reference points are not fixed minima and maxima, as participants can, throughout the experiment, give ratings below the initial minimum and above the initial maximum. The reference points serve like markers on a thermometer – a freezing point and a boiling point (therefore the name of the procedure).

Magnitude estimation was developed in psychophysics (Stevens 1946, 1951) and introduced to linguistics by Sorace (1992; see also Bard et al. 1996; Cowart 1997). In a magnitude estimation experiment, participants assign an arbitrary value (that has to be larger than zero) to a reference item and then rate each subsequent item in proportion to the reference item. Magnitude estimation is supposed to yield interval data (but see Sprouse 2011a for criticism). However, after an initial phase of excitement about this new technique for measuring linguistic acceptability (during which magnitude estimation was considered by some researchers as superior to other methods), it has now become "just" one method among others (and similar to other methods, magnitude estimation produces consistent results; cf. Weskott & Fanselow 2011 and Bader & Häussler 2010).

Experiment 3 examined 36 sentences that were randomly sampled from the same corpus that was used for Experiment 2 (see above). Of those 36 items, twelve sentences were unmarked in the original *LI* papers, twelve sentences were marked as unacceptable (*), and twelve sentences were marked as questionable (with marks such as ?, ??, and *?, and the like). Sentences were the same in all four sub-experiments.

3. Detecting non-cooperative behavior

Given that non-cooperative behavior affects the data beyond just creating zero-mean noise (for evidence see Section 4), we want to detect it and exclude non-cooperative participants. The literature on the challenges of crowdsourcing tasks used for research contains several suggestions as to how to identify suspicious responses (e.g. Kittur et al. 2008; Downs et al. 2010). However, such suggestions are typically rather task-specific. In the current section, we present a general detection tool based on response times as well as a task-specific tool for acceptability judgments. However, we start with a small typology of non-cooperative behavior.

3.1. A typology of non-cooperative behavior

We identify three main types of non-cooperative participants: simple spammers, clever spammers, and distracted participants. As those types of non-cooperative participants exhibit different behavior, we need different strategies to detect them.

Some participants click on response buttons (or type in ratings) without actually reading the sentences, let alone give meaningful ratings. This group of non-cooperative participants stands out by extremely fast response times. We call this type of non-cooperative participants "spammers", as they spam the data pool with nonsensical ratings (cf. Kazai et al. 2011). We distinguish two types of spammers: simple spammers (who just click their way through an experiment as fast as possible) and clever spammers (who pause every now and then to make their overall time look normal). Figure 1 and Table 2 serve to illustrate their behavior. A third group of non-cooperative participants reads the sentences and rates them, but their ratings do not bare out expected distinctions (i.e. the ratings appear "sloppy"). We label such participants "inattentive participants". And they are hard to identify: Their response times are normal and their ratings could in principle be "true".

However, when it comes to defining "inattentive", one has to strike a balance: We do not want to analyze output patterns and exclude every participant who does not fit the picture. In contrast to tasks for which *Amazon*'s *Mechanical Turk* was set up, for instance object identification on a photograph, transcribing audio recordings, etc., there is no answer that is correct a priori. We run acceptability judgments studies to find out how acceptable a certain construction is. But we can create items for which the status is uncontroversial. We call such items "booby trap" items and discuss them in more detail at the end of this section.

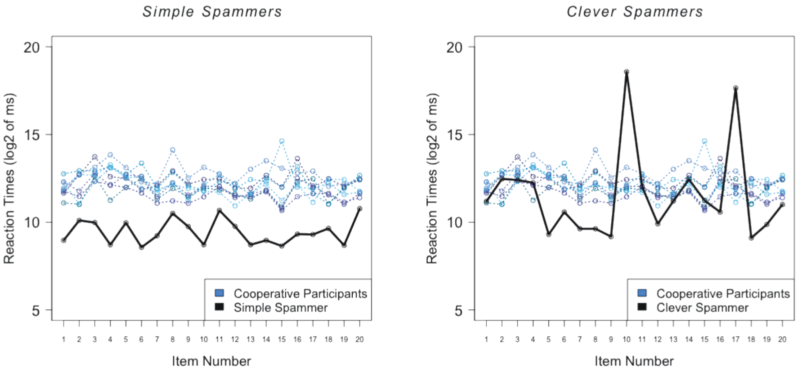


Figure 1. An illustration of the response times produced by cooperative participants (dashed gray lines) and two types of non-cooperative participants (solid black lines) – simple spammers (left) and clever spammers (right)

3.2. Analyzing response times

Spammers can be detected by analyzing response times. To detect simple spammers, a look at mean response times is sufficient. Clever spammers, however, make pauses which compensate for their very quick responses on the majority of items. As a result, a clever spammer's mean response time will be in the middle range of the sample or even prolonged (see Table 2). The median, however, which is robust to outliers, is expected to uncover clever spamming. To test this hypothesis and to determine the ratio of simple and clever spammers in our samples, we inspected both mean and median response times.

Mean response times are computed by dividing the sum of all response times of this participant (x_1, x_2, ... x_n) by the number of items (n). Determining the median involves two steps. (i) Response times for each participant are ordered from shortest to longest, (ii) the middle value is picked out. In experiments like ours with an even number of items, the middle value is the mean of the two middlemost response times. Experiment 2, for instance, included 108 items (i.e. each participant contributed 108 response times). The median in this example is the average of the 54th and 55th response time when ordered numerically.

For purposes of illustration, we discuss the data from Part 1 of Experiment 2a in more detail. Table 2 provides selected (rounded) response times by the first seven participants. Items 1–3 are from the set of sentences left unmarked in the corresponding *LI*-papers, items 4–6 are marked as ungrammatical (*) in the original papers.[6] Also, the response times include time for reading: The onset for measuring the response time is the onset of displaying the item.

Participant 7 stands out as being extremely fast, as his/her mean response time is way below the mean response times of the other subjects. Participant 6 is a clever spammer: Due to the long pause on Item 5, the mean response time is rather long (this also holds when all data points contributed by this participant are taken into account; mean response time: 7,994 ms, compared to 3,947 ms for all participants; interquartile range for all participants: [3,960, 6,568]). However, as soon as we look at median response times, both spammers stand out. A critical question is: How extreme is too extreme? Participant 1, for instance, is rather fast compared to Participants 2–5, though not as fast as Participants 6 and 7. Should we exclude Participant 1?

Table 2. Selected response times (in ms) for the first seven participants in Experiment 2a, Part 1

	Item							
Participant	**1**	**2**	**3**	**4**	**5**	**6**	**Mean**	**Median**
1	1,710	1,890	5,090	2,310	2,980	2,070	2,675	2,190
2	2,210	2,800	11,280	7,030	4,260	3,200	5,130	3,730
3	3,340	2,170	4,470	3,060	2,790	3,020	3,192	3,040
4	3,390	5,050	8,040	5,460	5,520	5,720	5,530	5,490
5	1,900	1,920	5,820	4,710	3,460	5,560	3,895	4,085
6	750	460	490	740	**390,850**	550	**65,640**	**645**
7	380	460	990	870	420	800	**653**	**630**

[6] Rejecting a sentence typically takes a bit longer, therefore the trend for longer response times for items 4–6. However, the two sets (originally unmarked versus *-marked items) are not matched in terms of length.

The problem is related to outlier detection in experiments collecting any type of response times. Unfortunately, the literature in psychology and psycholinguistics is more concerned with extremely long response times than with unrealistically short ones. Ratcliff (1993) argues that short response times by non-cooperative participants are easy to spot and disregard; and thus, no explicit detection strategies are needed. However, based on our own experience and data, we find it not that easy. But note that Ratcliff and others are concerned with individual data points while we are concerned with participants.

There is no gold standard for dealing with outliers when it comes to response times (see Miller 1991 and Cousineau & Charter 2010 for overviews). Some researchers set absolute cut-off points (typically for reading or fixation times in self-paced reading or eye-tracking experiments), while others use relative thresholds, typically based on standard deviations or the interquartile range (for linguistics, see Baayen 2008 and Gries 2013; for a detailed discussion see Rousseeuw & Croux 1993). A survey in the field of psychology (Leys et al. 2013) shows that most studies use standard-deviation approaches.

In standard-deviation approaches, response times that fall within a given number of standard deviations from the mean are considered as outliers. The rationale behind this criterion is related to the characteristics of the normal distribution. In a normal distribution, 99.7% of the data are located in the range of the mean ±3 standard deviations. Everything outside this range can be considered as extreme outliers. Less conservative thresholds are 2.5 or even 2 standard deviations below/above the mean. While the application of these values works for identifying single extreme data points in big samples, it fails for the identification of non-cooperative participants, especially if there are many of them. The standard deviation measure used to detect outliers, is itself affected by outliers. Non-cooperative participants contribute extremely low median response times and thereby inflate the standard deviation.[7] Take for instance the sample in Table 2. Since we are looking for suspicious participants, we first computed median response times for each

[7] Below, we present the MAD, a more robust method that is well suited to detect outliers.

participant (last column in Table 2) and then the mean of those medians. The mean of median response times in this subset is 2,830 ms with a standard deviation of 1,804 ms. Subtracting two standards deviations or more gives an absurd threshold for the lower boundary: a negative value.

We therefore chose a less conservative criterion of 1.5 standard deviations for the lower boundary. A 1.5 standard-deviations criterion gives a slightly broader interval than the interquartile range, which in normally distributed data corresponds to ±1.35 standard deviations around the mean. In a normal distribution, a range of 1.5 standard deviations around the mean covers around 87% of the data. Given the proportion of non-cooperative participants reported in the literature, this seems a good range to look at. The lower threshold (θ_{lower}) for excluding participants as non-cooperative is calculated using the formula in (7), where n is the number of participants, x_i is the mean or median response time for participant i and $\bar{x}$ is the grand mean response time averaged over the means or medians of all participants.

$$(7) \qquad \theta_{\text{lower}} = \frac{1}{n}\sum_{i=1}^{n} x_i - 1.5\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2}$$

For the sake of comparability, we did the same for the mean of means. Table 3 gives the resulting thresholds for Experiment 2a, Part 1. When applied to mean response times, the 1.5 standard-deviation criterion identifies the simple spammer (Participant 7) but misses the clever spammer (Participant 6). Applied to median response times, both spammers are marked as outliers.

Table 3. Lower thresholds for part 1 of Experiment 2a, applying different methods for detecting spammers

Measure	**Mean**	**SD**	**θ_{lower}**	**Participant 6**	**Participant 7**
Mean RTs	5,467	2,266	2,068	7,749	883
Median RTs	3,974	1,418	1,848	1,490	745

So far, we only looked at extremely short response times and set a lower boundary. Do we need an upper boundary as well? Participant 4 in Table 2 has a comparatively long median response time. Shall we exclude this participant? Does the long median response time indicate a form of non-cooperative behavior (e.g. repeated switching between the linguistic task and some other activity)? Or perhaps Participant 4 is simply a slow reader and a very careful rater?

Three possible sources for long response times need to be distinguished: (i) slow reading and judging (which is fine and results in overall longer response times, including the mean; this seems to be the case for Participant 4), (ii) single disruptions (which lead to increased response times for single items but do not affect the median response times), and (iii) frequent disruptions throughout the entire experiment (which increase the median response times). We want to exclude the last type, because frequent distractions arguably reduce the quality of responses.

Applying a 1.5 standard-deviation criterion to extremely long response times yields an upper boundary of about 6 seconds in Experiment 2. Six seconds per sentence is not overly slow and in our view, there are no reasons to exclude a participant with such reaction times. In contrast to standard procedures for outlier detection, we therefore suggest an asymmetric exclusion criterion: a lower criterion at mean median reaction time -1.5 standard deviations (about two seconds for our experimental stimuli), but an upper criterion at mean median reaction time +4 standard deviations (about 10 seconds).

For the first part of Experiment 2a, we observed that the analysis of median response times is more successful in identifying the type of non-cooperative participants which we call spammers. As mentioned above, we distinguish two types – simple spammers with extremely fast response times throughout and clever spammers with very many extremely fast response times counterbalanced by a few very long response times. Crucially, the mean-based criterion detected only one of two spammers whereas a median-based criterion detects both spammers.

For the entire data set, the mean-based criterion detected only 11 participants with non-cooperative behavior compared to 25 participants detected by a median-based criterion (see Table 4). In other words, 15 spammers adjusted their mean response times by pausing. Apparently, clever spammers are aware that their response times can be tracked.

Table 4. Distribution of non-cooperative participants in the three experiments

Experiment	Simple spammers	Clever spammers	Inattentive participants
1	7	8	5
2	2	3	4
3	2	4	1
Total	**11**	**15**	**10**

Though the median-based approach is more robust than a mean-based approach, the application of a standard-deviation criterion is vulnerable to extreme values. Note that spammers do not only contribute extreme individual response times but also extreme median response times. Their median response times decrease the mean of median response times and increase the standard deviation for median response times. Thus, a criterion based on the median of the median response times might be an even better estimator. The median absolute deviation (MAD)[8] provides a robust measure of dispersion (cf. Hampel 1968, 1974; Huber 1981) and is therefore well suited to detect outliers. The calculation of the MAD requires the following steps (summarized in the formula in (8) taken from Huber 1981: 107): (i) determining the median *Mj* for the series of observations, (ii) determining the absolute deviations from this median by subtracting from each value in the vector and taking the absolute of the result, and (iii) determining the median *Mi* of the series of absolute deviations.

(8) $$MAD = M_i\left(\left|x_i - M_j(x_j)\right|\right)$$

We tested an MAD-approach to a subset of our data set (Experiment 2a). Following the recommendation of Leys et al. (2013), we adjusted the MAD by a factor 1.4826 and used a decision criterion of 2.5. Thus, par-

[8] The median absolute deviation (from the median) should not be confused with the mean absolute deviation, which is also abbreviated as MAD.

ticipants with a median response time that is 2.5 x 1.4826 MAD below the median of the median response times were rejected.[9] For Part 1, the resulting lower boundary is pretty close to the boundary in our 1.5 standard-deviation approach (1,825 vs. 1,848 ms); for Part 2, however, the MAD-criterion is too conservative (892 vs. 1,990 ms) and misses both spammers. We therefore stick to the standard-deviation approach and leave modifications of an MAD-approach to future research.

As a final remark, note that using outlier-detection procedures for identifying non-cooperative participants with extremely fast response times and distracted participants with extremely long response times only works as long as they are outliers in the sense of being exceptional. Methods for outlier detection rely on the assumption that the underlying distribution is unimodal and symmetrical but contaminated by outliers producing a heavy tail. Even a median-based approach will not be successful if a large number of the participants were non-cooperative. The median's breakdown point is at 50% (which is the highest breakdown point possible).[10] When non-cooperative participants make up half or more of the sample, we need an absolute criterion for what is considered to be a realistic response time or an independent criterion, for example the performance on what we call "booby trap items".

3.3. Booby trap items

Analyzing response times, more specifically looking for extreme median response times, allows for the detection of simple and clever spammers. Inattentive participants, however, have average response times and can therefore not be detected by response-time based criteria. To detect inattentive participants, we need to examine the actual ratings. Yet we do not want to exclude every participant who does not fit the picture. Instead, the

[9] The MAD needs to be adjusted to be consistent at the underlying distribution. For a normal distribution, the MAD must be divided by 0.6745 (Huber 1981: 108), which is equivalent to multiplying it by 1.4826.

[10] The breakdown point gives the fraction of contaminated observations that an estimator can cope with before giving incorrect results. The mean and the standard deviation have a breakdown point of 0 since they can be contaminated by a single observation. The interquartile range breaks down at 25% (Huber 1981).

selection needs to be guided by objective criteria. We suggest basing such a selection on actual ratings by inserting items that are specifically designed for filtering out suspicious participants. We call such items "booby trap items", as they are non-critical items whose status is well-established (e.g. clearly acceptable or clearly unacceptable). Participants who fail to distinguish "bad" and "good" items should be excluded from analyses.[11]

In the following, we give an overview of the use of booby trap items in our three experiments. Experiment 1 did not include any booby trap items. For Experiment 2, we created (9) and (10) as good items representing constructions that are acceptable in (North) American English but are fairly marked in other varieties,[12] and (11) and (12) as bad items being instances of constructions that are acceptable in Indian English but not in (North) American English (Hansen et al. 1996). The items were randomly interspersed in the last two thirds of the questionnaires.

(9) My son's grades have gotten better since he moved out of the fraternity.
(10) The professor requested that Dillon submit his research paper before the end of the month.
(11) Peter wanted that we should come early.
(12) My knowledges of chemistry are rather weak.

In our analysis, we used the following criterion: If the two marked booby trap items received a higher average rating than the two good ones, then we excluded that participant. Based on this criterion, we identified four inattentive participants in Experiment 2 (two of them had a median response time close to the lower boundary).

Experiment 3 did not include genuine booby trap items but four items that were designed for calibration purposes (two bad ones and two good ones). As the status of the calibration items was determined in a previous experiment, we (post-hoc) analyzed the calibration items in

[11] Booby trap items can also be used to filter out speakers who are not speakers of the variety under investigation, e.g. American English versus Indian English, etc. For this, one needs to know which variety (dis)allows certain types of constructions.

[12] The items are modelled after Brians (2013) and Kövecses (2000), respectively. Ideally, such booby trap items should be clearly marked, but this was hard to achieve for our purposes.

Experiment 3 in a similar way to our analysis of booby trap items.[13] Sentences (13) and (14) serve to fix the middle section of the scale. Sentences (15) and (16) do the same for the middle part of the scale; they come from Ferreira & Swets (2005) and received mediocre to bad ratings in another experiment of one of the authors (Juzek 2016).

(13) As Obama's top counterterrorism adviser, Brennan has helped manage the drone program.
(14) Iran has proposed restarting talks as early as next month.
(15) This is a donkey that I don't know where it lives.
(16) This is the man that I don't know where he comes from.

While the use of such "booby trap" items is common practice in psycholinguistics, we do not know of many judgment studies in linguistics that make use of them. However, we strongly recommend their use, as in our experiments, about 4% of the participants failed on these items (for absolute numbers see Table 4). Yet, we do not recommend relying on booby trap items only since several of our spammers managed to pass this test. They accidentally pressed the right button. However, the percentage of spammers that are correct by chance probably decreases with a higher number of booby trap items.

4. The effect of non-cooperative behavior

Previous research in other domains than linguistics has shown that non-cooperative behavior affects the data in serious ways (e.g. Downs et al. 2010; Kazai et al. 2011; Eickhoff & de Vries 2013). Our study aims to assess the impact of non-cooperative behavior in linguistic judgment studies.

[13] A note on calibration items: Faced with any given scale (in case of Experiment 3b, a 5-point scale), each participant will define the points on the scale slightly differently. Participant A might interpret "5" as "OK" while Participant B interprets "5" as "truly exceptional" or "stylistically brilliant". Likewise, interpretations of the other endpoint of the scale, as well as of the points that lie in-between, will vary between participants. The first few items in a questionnaire play a decisive role for such scale biases. They serve as anchor points for subsequent items. Calibration items deliberately chosen for anchoring and presented at the beginning of a questionnaire might mitigate the effect of scale biases.

Applying the detection criteria discussed above, we identified non-cooperative behavior in eleven of the twelve experimental sessions of our study. For each session, we compared the overall mean ratings by the cooperative participants to the overall mean ratings by the non-cooperative participants. The results are shown in Table 5. Wilcoxon Signed Rank Tests indicate that in eight sessions the means differ significantly ($p < 0.05$). The finding that non-cooperative participants contribute non-zero means noise underscores the need to detect them and exclude them from analysis.

The easiest way to click one's way through an experiment is to click repeatedly on the same button. Under this assumption, ratings by non-cooperative participants are expected to show less variation compared to cooperative participants. To test this hypothesis, we calculated the variance within each participant. Table 5 gives the averaged individual variances separately for the group of cooperative participants and non-cooperative participants. The pattern mirrors the findings for the overall means. Again, in eight sessions we find a difference between cooperative and non-cooperative participants. As expected, ratings by non-cooperative participants show less variation compared to ratings by cooperative participants; and the direction of the effect is the same in all cases. We take this finding as a clear indicator that ratings by non-cooperative participants are less informative and inferior in quality.

5. Strategies against non-cooperative behavior

Detecting and excluding non-cooperative participants is a means of reducing non-cooperative behavior post-experiment. However, there are also ways to reduce non-cooperative behavior both before and during the experiment. Section 5.1 discusses filtering as a pre-study prevention technique recommended by Eickhoff & de Vries (2013) and others. Section 5.2 introduces a warning mechanism that effectively stops participants from clicking through an experiment.

Table 5. Comparison of cooperative and non-cooperative participants across our twelve sessions: (i) grand means (with standard deviations in parentheses), including the results of Wilcoxon Signed Rank Tests, and (ii) the mean variance, including results of the F-Test

Experiment			Scale		Grand Mean (SD)			Variance		
					Cooperative	Non-coop.	*p*	Cooperative	Non-coop.	*p*
1	a	Auditory	5pt. Likert scale		3.61 (1.47)	3.74 (0.85)	*ns*	1.41	0.83	*
		Visual			3.54 (1.48)	3.24 (1.23)	*	1.44	1.03	*
	b	Auditory	5pt. Likert scale		3.63 (1.47)	3.34 (0.81)	*	1.40	0.80	*
		Visual			3.64 (1.46)	3.54 (1.17)	*	1.38	1.15	*
2	a	Visual	7pt. Likert scale	Part 1	0.54 (0.33)	0.63 (0.32)	*	0.33	0.31	*ns*
				Part 2	0.49 (0.33)			0.32		
	b	Visual	Binary	Part 1	4.41 (2.11)	4.46 (1.71)	*ns*	2.03	1.64	*
				Part 2	4.28 (2.12)	3.77 (1.50)	*	2.04	1.44	*
3	a	Visual	Binary		0.49 (0.50)	0.67 (0.48)	*	0.49	0.48	*ns*
	b	Visual	7pt. Likert scale		4.07 (2.19)	3.76 (1.40)	*ns*	2.01	1.19	*
	c	Visual	Thermometer judgment		0.48 (0.34)	0.61 (0.16)	*	0.30	0.14	*
	c	Visual	Magnitude estimation		1.33 (0.79)	1.10 (0.75)	*	0.63	0.51	*ns*

Note. * $p < .05$; *ns* = not significant ($p > .05$)

5.1. Banning non-cooperative participants

The group of non-cooperative participants has a specific demographic. The typical non-cooperative participant is a young man in his twenties (Downs et al. 2010). Hence, one way to reduce the proportion of non-cooperative participants would be to exclude young men. However, this comes at the cost of a non-representative sample and excludes potentially cooperative participants. In our study, we therefore pursued a different strategy. Following recommendations by Eickhoff & de Vries (2013), we used filtering by prior performance. In a *Mechanical Turk* study, participants ("workers" in *Mechanical Turk* terminology) are evaluated after an experiment. We used the participants' reputation as an entry criterion for the study. Participants had to have an approval rate of at least 98%. In addition, we used experience as an entry criterion, operationalized as the number of approved *Mechanical Turk*-tasks ("HITs"). Participants had to have a minimum of 5,000 HITs.[14]

As evidenced by the number of non-cooperative participants in our study, setting entry criteria does not prevent non-cooperative behavior completely. Furthermore, banning potentially non-cooperative participants comes with a trade-off. Using lenient criteria increases the risk of including a considerable proportion of non-cooperative participants. Using very strict criteria, on the other hand, increases the risk of a sampling bias (i.e. the group of participants will be rather homogenous and most likely not representative). We suggest using rather lenient entry criteria, but accompany them by prevention techniques applicable during the course of an experiment. We discuss those in turn.

5.2. Discouraging non-cooperative behavior

Despite banning, some potentially non-cooperative participants might still find their way into one's study, especially if we refrain from very strict pre-screening criteria. It is therefore useful to discourage participants from acting non-cooperatively. To do so, Experiments 1 and 2

[14] Note that these criteria are less strict than what is required to become a *Mechanical Turk* "master worker".

included an on-line warning mechanism that produced an alerting pop-up window when a participant's response times were extremely short. In the sub-experiments with binary judgments (Experiment 1a) or Likert-scales (Experiments 1b, 2a, and 2b), the alert popped up when response times repeatedly fell below 400 ms. This threshold is related to a formula presented in Bader & Häussler (2010) for deriving presentation times per word in a speeded-grammaticality-judgments experiment. For the warning mechanism, we used a baseline per sentences and added additional time per character to compensate for length differences. The formula for determining the response time threshold is given in (17), where *lc* is the sentence length in characters. We lowered the critical reading time even further by dividing the resulting sum by two.

(17) $\theta_{\text{warning}} = (225 + 25 \times l_C)/2$

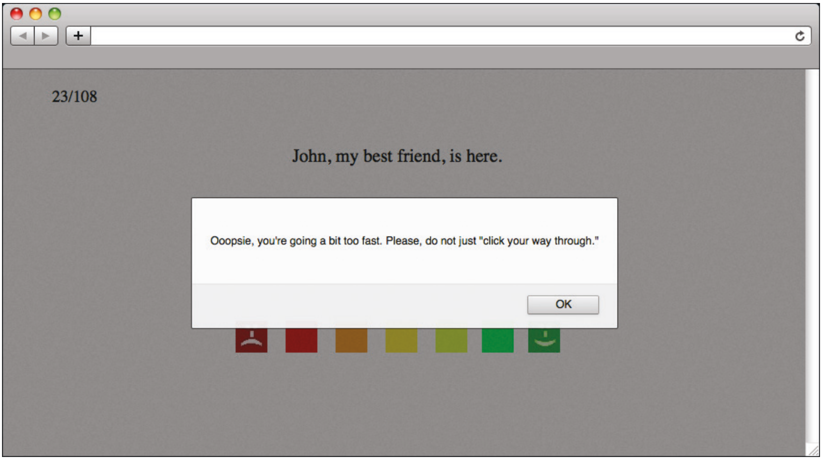
For the shortest sentence in Experiment 2, the formula outputs a critical response time θ_{warning} of 400 ms. Response times below that threshold are unrealistically fast. We consider it impossible to read and judge a sentence that fast, even for fast readers.[15]

The threshold was designed with a Likert-scale in mind. Thermometer judgments and magnitude estimation require participants to calculate numerical values, click on a text box, type in the value, and click on a button to submit the rating. To adjust for these extra-processes, we added another 800 ms and defined a threshold of 1,200 ms. Note that this is still extremely fast (and way below the mean response times), even in an experiment that does not involve those extra-steps.

In Experiment 2 the first alert appeared when the response time was below the threshold for the fourth time. The warning message appeared at the middle of the screen and was rather friendly: "Ooopsie, you're going a bit too fast. Please, do not just 'click your way through'." (see

[15] For comparison: Subtitles are typically displayed for one to six seconds. Longer displays lead to rereading, shorter displays make it difficult to follow. The six-seconds standard, which is widely followed for dubbing, states that two lines containing 32–39 characters each should be displayed for six seconds. This corresponds to a reading speed of 11–13 characters per second or 130–150 words per minute (for an overview of technical aspects of dubbing and in particular the temporal dimension see Diaz Cintas & Remael 2007).

Figure 2). If the participant continued fast clicking, a second alert occurred that was less friendly and announced consequences: “Sorry, you’re going too fast and you might not get approved. If you are getting this message although you’re doing the task properly, please continue as before.” In Experiment 3, the alerts were the same and occurred after the same fraction of response times violating the threshold.



23/108

John, my best friend, is here.

Ooopsie, you're going a bit too fast. Please, do not just "click your way through."

Figure 2. An illustration of the (first) warning message

The 400 ms threshold is very low, even spammers only occasionally come below this threshold.16 And yet, the warning mechanism turned out to be effective. Experiment 1 did not make use of the warning mechanism and we had to exclude 15 out of 142 participants (11%) for being too fast. The rate of participants with response times being too fast was much lower in the other two experiments that included the warning mechanism: 5 out of 143 participants in Experiment 2 (3%) and 6 out of 137 participants in Experiment 3 (4%). Thus, including the response-

[16] In our study, typical response times (including time for reading) by cooperative participants are between one and nine seconds. Experiment 2a, for instance, has a mean response time of about 5,500 ms, the median is roughly 4,000 ms, exceeding the 400 ms threshold by a factor of 10.

time based warning mechanism reduced this kind of non-cooperative behavior by more than 50%. Apparently, participants stop being blatantly non-cooperative once they realize that (clearly) non-cooperative behavior can be detected.

6. Conclusions

Many researchers have raised concerns about data quality in acceptability judgments (and other types of data) collected online. As shown, non-cooperative behavior has a significant negative impact on the quality of one's data. However, only few studies apply adequate measures to prevent, discourage and detect non-cooperative behavior. In the present paper, we discussed relevant strategies and their effectiveness.

Despite strong banning, non-cooperative participants made up about 14% of the sampling population in our experiments. They come in different types and require different means to be detected. Spammers, that is, participants submitting unrealistically fast responses, stand out when screening response times. Based on our data, we recommend exclusion criteria based on median response times. While mean-based response time criteria can be used to detect simple clicking-through, only median-based criteria can detect clever spamming. We suggest either a standard-deviation approach with an asymmetric decision criterion (1.5 standard deviations below and 4 standard deviations above the grand mean averaged over medians) or an MAD-approach as outlined at the end of Section 3.2.

In addition to detecting non-cooperative behavior, response times can be used to discourage non-cooperative behavior in the first place. A warning mechanism against extreme response times reduced the proportion of non-cooperative behavior in our studies by about 50%. In the present study, the threshold for the warning mechanism was rather extreme – 400 ms while most response times were longer than 1,000 ms. Further studies are required to determine the optimal threshold.

Finally, inattentive participants, a third group of non-cooperative participants, could be caught by booby trap items. When carefully selected, booby trap items can be an effective detection strategy.

In summary, our results lead us to the following four recommendations for acceptability judgments studies:

1. Apply some form of pre-screening of participants (e.g. through approval rates or task experience).
2. Include booby trap items and use them for post-experiment exclusion of participants who fail on these items for whatever reason (e.g. distraction or cross-linguistic differences).
3. Collect and track response times. Use them online to give a warning when response times fall below a predefined extreme threshold.
4. Screen response times post-experiment. Look for extreme median response times.

While implementing these strategies requires slight adjustments in experimental design and procedure, it is worth the effort, as the quality of one's data improves significantly.

References

Baayen, Harald. 2008. *Analyzing linguistic data*. Cambridge: Cambridge University Press.

Bader, Markus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46. 273–330.

Bard, Ellen Gurman, Dan Robertson & Antonella Sorace 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1). 32–68.

Becker, Misha Karen. 2006. There began to be a learnability puzzle. *Linguistic Inquiry* 37(3). 441–456.

Brians, Paul. 2013. *Common errors in English usage*. Washington: William, James & Co.

Cousineau, Denis & Sylvain Chartier. 2010. Outliers detection and treatment: A review. *International Journal of Psychological Research* 3(1). 58–67.

Cowart, Wayne 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks: Sage.

Dandurand, Frederic, Thomas Shultz & Kristine Onishi. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods* 40(2). 428–434.

Diaz Cintas, Jorge & Aline Remael. 2007. *Audiovisual translation: Subtitling.* Manchester: St. Jerome Publishing.

Downs, July S., Mandy B. Holbrook, Steve Sheng & Lorrie Faith Cranor. 2010. Are your participants gaming the system? Screening Mechanical Turk workers. In Elizabeth Mynatt, John White & Gerrit van der Veer (eds.), *CHI 2010: Proceedings of the SIGCHI conference on human factors in computing systems,* 2399–2402. Atlanta: Association for Computing Machinery.

Eickhoff, Carsten & Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval* 16(2). 121–137.

Featherston, Sam. 2008. Thermometer judgments as linguistic evidence. In Riehl, Claudia Maria & Astrid Rothe (eds.), *Was ist linguistische Evidenz?,* 69–90. Aachen: Shaker.

Ferreira, Fernanda & Benjamin Swets. 2005. The production and comprehension of resumptive pronouns in relative clause "island" contexts. In Ann Cutler (ed.), *Twenty-first century psycholinguistics: Four cornerstones,* 263–278. Mahwah: Erlbaum.

Gibson, Edward, Steve Piantadosi & Kristina Fedorenko. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass* 5. 509–524.

Gries, Stefan. 2013. *Statistics for linguistics with R.* Berlin: Mouton de Gruyter.

Hampel, Frank. 1968. *Contributions to the theory of robust estimation.* Berkeley: University of California dissertation.

Hampel, Frank. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69. 383–393.

Hansen, Klaus, Uwe Carls & Peter Lucko. 1996. *Die Differenzierung des Englischen in nationale Varianten: Eine Einführung.* Berlin: Erich Schmidt.

Häussler, Jana & Tom Juzek. 2015. Lab or Armchair? The benefits of formal acceptability judgements. *Chicago Linguistic Society (CLS)* 51.

Huber, Peter J. 1981. *Robust statistics.* New York: Wiley.

Juzek, Tom S. 2016. *Acceptability judgement tasks and grammatical theory.* Oxford: Oxford University dissertation.

Kazai, Gabriella, Jaap Kamps & Natasa Milic-Frayling. 2011. Worker types and personality traits in crowdsourcing relevance labels. In Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis & Ian Ruthven (eds), *Proceedings of the 20th ACM international conference on information and knowledge management (CIKM 2011),* 1941–1944. New York: Association for Computing Machinery.

Kittur, Aniket, Ed H. Chi & Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In Margaret Burnet, Maria Francesca Costabile, Tiziana Catarci, Boris de Ruyter, Desney Tan, Mary Czerwinski & Arnie Lund (eds.), *CHI 2008: Proceedings of the SIGCHI conference on human factors in computing systems,* 453–356. New York: Association for Computing Machinery.

Kövecses, Zoltan. 2000. *American English – an introduction.* Peterborough: Broadview Press.

Krantz, John H. & Reeshad Dalal. 2000. Validity of web-based psychological research. In Michael H. Birnbaum (ed.), *Psychological experiments on the internet,* 35–60. New York: Academic Press.

Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard & Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49(4). 764–766.

Martin, Roger. 2001. Null Case and the distribution of PRO. *Linguistic Inquiry* 32(1). 141–166.

Miller, Jeff. 1991. Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology* 43(4). 907 –912.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In Chris Callison-Burch & Mark Dredze (eds.), *CSLDAMT 2010: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk,* 122–130. Stroudsburg: Association for Computational Linguistics.

Newmeyer, Frederick. 2003. Grammar is grammar and usage is usage. *Language* 79. 682–707.

Phillips, Colin. 2010. Should we impeach armchair linguists? In Shoishi Iwasaki, Hajime Hoji, Patricia M. Clancy & Sung-Ock Sohn (eds.), *Japanese/Korean Linguistics,* vol. 17, 49–64. Stanford: CSLI Publications.

Ratcliff, Roger 1993. Methods for dealing with reaction time outliers. *Psychological Bulletin* 114. 510–532.

Rousseeuw, Peter J. & Christophe Croux 1993. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association* 88(424). 1273–1283.

Schnoebelen, Tyler & Victor Kuperman. 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija* 43(4). 441–464.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press

Sorace, Antonella. 1992. *Lexical conditions on syntactic knowledge: Auxiliary selection in native and non-native grammars of Italian*. Edinburgh: University of Edinburgh dissertation.

Sprouse, Jon. 2011a. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87(2). 274–288.

Sprouse, Jon. 2011b. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134. 219–248.

Stevens, Stanley Smith. 1946. On the theory of scales of measurement. *Science* 103. 677–680.

Stevens, Stanley Smith. 1951. Mathematics, measurement and psychophysics. In Stanley Smith Stevens (ed.), *Handbook of experimental psychology*, 1–49. New York: Wiley.

Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115. 1481–1496.

Weskott, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273.

The dot plot: A graphical tool for data analysis and presentation

Lukas Sönning
University of Bamberg

Cleveland's (1984) introduction of the dot plot to the scientific community dates back more than 30 years. Its clarity, flexibility, and efficiency make it a useful tool that is applicable to a wide range of descriptive and inferential analyses. Yet, this graph type has not gained the currency it deserves; in fact, it appears to be unknown to most researchers (Jacoby 2006; Keen 2010). This paper presents the dot plot and brings together various extensions that have emerged over the last 30 years. Advantages over alternative chart types are illustrated and design options and recommendations for the display of more complex data sets are discussed. The application of dot plots to quantitative data in linguistics is demonstrated, focusing on examples from corpus linguistics, meta-analysis and statistical modeling. The final sections reflect on important limitations of this display type and refer the reader to software for the implementation of dot plots. An online appendix provides a brief R tutorial as well as templates for Microsoft Excel, which allow for easy production of dot plots by entering data into spreadsheet templates.

1. Introduction

Graphs are indispensable tools in quantitative research since they reveal structure in the data in an effective and accessible way. A functional distinction is often made between graphs for data analysis and data presentation (Fienberg 1979; Schmid 1983). Graphing in data analysis serves to communicate between researcher and data. It is an iterative process and involves drawing many displays to gain different perspectives on a data set (Unwin 2015). Presentation graphs, on the other hand, aim to effectively communicate findings to an audience. To this end, principles of visual perception should guide the choice of graph type and graphical parameter settings to obtain an effective display.

This paper introduces the dot plot (Cleveland 1984), a display method suitable for both data analysis and presentation. It is an (unjustly) underutilized graph type that appears to be unfamiliar to most researchers (Jacoby 2006; Keen 2010). Its conceptual simplicity, however, makes it a versatile tool for many types of statistical analyses. The design of the dot plot is inspired by insights gained from research on visual perception, the aim being an optimization of the decoding of quantitative information. There are also several practical advantages compared to other more widely used chart types, such as the bar chart. It is the aim of this paper to demonstrate the usefulness and added value of the dot plot and argue for its routine usage in quantitative research (for examples of their application in linguistic research see Werner & Fuchs 2016; Krug et al. 2016; Schützler forthcoming).

After an outline of the theoretical background on graphs in scientific research, Section 3 introduces the simple dot plot, including the relevant terminology and a number of extensions for more complex data sets. Next, advantages over alternative chart types are summarized and illustrated. Section 5 discusses design options and gives recommendations on the construction of dot plots. Applications to linguistic data analysis are demonstrated in Section 6, including usage in simple meta-analyses and in the investigation of binary and frequency outcomes in corpus linguistics. The final sections reflect on the limitations of dot plots and discuss their implementation in R and Microsoft Excel. An online appendix includes brief tutorials for dot plots in R and spreadsheets for their implementation in Excel.

2. Theoretical background

The discussion and comparative evaluation of graph types can build on theoretical insights gained across a wide range of disciplines. These include exploratory data analysis (Tukey 1977), experimental research on graphical perception (Cleveland 1993), psychology (Wertheimer 1938) and neuroscience (Kosslyn 2006). This section aims to lay a conceptual and terminological foundation and elaborates on four aspects: (i) the purpose of statistical graphs, (ii) the active process of decoding information from a graph, (iii) a model of graphical perception, and (iv) psychological principles of graph perception and design. Key terms are italicized throughout the paper.

The purpose of graphs

Tukey (1993: 2) concisely states the “true” purpose of graphs: first, graphs are not meant to communicate precise values, but are rather *semi-quantitative*; exact numbers should be provided in tables. Second, graphs are for *comparisons*. As pointed out by Tufte, “at the heart of quantitative reasoning is a single question: *Compared to what?*” (1990: 67, emphasis in original). Third, graphs are for *impact* on the viewer – important information must be easily discernible. In short, the purpose of a graph is to “force” the viewer to make key *comparisons* of interest in a *semi-quantitative* manner. According to Tukey (1993: 3), such *semi-quantitative comparisons* yield statements like “is way above”, “is above”, “is a little above”, “is almost equal to/is almost on”, “is a little below”, “is below”, “is way below”.

Decoding information from a graph

In order for such *semi-quantitative comparisons* to be made, the viewer must formulate a conceptual question, a piece of information to be extracted from the graph (Pinker 1990: 94). In other words, not every piece of information can be forced upon the viewer; rather, he or she plays an active role in decoding information from a display. This operation can be conceived of as a two-step process (Ware 2013: 139). First, a *visual query* is formulated, which identifies the problem to be solved or question to be answered. The second step is the *visual search*, the decoding of the display in response to the query, whereby the viewer identifies relevant patterns in the display. The success of a visual display thus also depends on the viewer (and data analyst), who must know where to look and what to look for.

A model for graphical perception

The *visual search* is an active process guided by principles of visual perception. Based on experimental research, Cleveland (1993) proposed a model for graphical perception. It introduces a number of useful terms for the description of displays and the mental operations involved in decoding information. Graphs encode *quantitative* and/or *categorical* variables. Quantitative variables yield values or measurements; categorical variables (binary, nominal and ordinal) assign observations to different groups or categories. The displayed content of a graph can be divided into *physical* and *scale* information. *Physical* information refers to the ink (or pixels) shown, excluding numeric and category labels (i.e. numbers on the axes and labels in the key). Such labels provide *scale* information and assign numbers (in the case of quantitative variables) and labels (for categorical variables) to the *physical* information in the display. According to Cleveland's (1993) model, graphical perception involves two mental operations: (i) *pattern perception*, which refers to the decoding of *physical* information, and (ii) *table look-up*, which refers to the decoding of *scale* information. *Pattern perception* in turn involves three visual operations: (i) *detection*, the recognition of physical elements, (ii) *assembly*, the grouping of elements belonging to the same category, and (iii) *estimation*, the *comparison* of visual elements.

Psychological principles

Pattern perception is governed by general principles of cognition; these help explain how humans decode visual information and thus inform graph construction. Kosslyn (2006) formulates eight psychological principles of effective graph design. These include the audience-oriented principles of *relevance* (show only relevant information) and *appropriate knowledge* (take into consideration the prior knowledge of the audience). Concerned with the visual appearance of the graph are the principles of *salience* (prominent elements receive more attention), *discriminability* (elements have to be sufficiently different to be distinguishable) and *perceptual organization*. The last set of Kosslyn's principles focuses on communication and includes the principles of *compatibility* (form must match content), *informative changes* (changes in form must signal changes in content) and *capacity limitations* (do not overload your audience's working memory). Of particular importance is the principle of perceptual organization, which includes the notion of pre-attentive attributes of stimuli, which affect *detection* and discriminability, and Gestalt laws of perception (Wertheimer 1938; Ware 2013: 181–199). The latter facilitate *assembly* – that is, the selective perception of entities belonging to the same group.

Gestalt laws include the law of *similarity* (similar elements will be grouped together), *proximity* (close elements will be grouped), *good form* (regular or symmetric shapes are perceived as single units) and *connectedness* (linked elements will be grouped; Palmer & Rock 1994).

Theoretical insights into graph design and perception provide a useful foundation for the informed application of statistical graphs in quantitative research. As such, they can guide the choice between different graph types and design options for the display of a particular data set.

3. The dot plot

The dot plot was introduced by Cleveland (1984) as a graphical display of labeled data. Figure 1 shows a simple dot plot of the relative frequency of the 10 most frequent nouns in the *British National Corpus* (BNC; Leech et al. 2001). The horizontal scale encodes a quantitative variable (frequency), the vertical scale a categorical variable (noun); light horizontal lines connect the data points with their labels. Labeled data – that is, numeric values with labels – are common in data analysis. They occur in the form of raw data (e.g. individual measurements or counts in a corpus), summary statistics (e.g. measures of central tendency/location and dispersion/ spread, percentages or other effect sizes) and model parameters (e.g. regression coefficients and information criteria). Dot plots can therefore be put to use in a wide range of descriptive and inferential analyses.

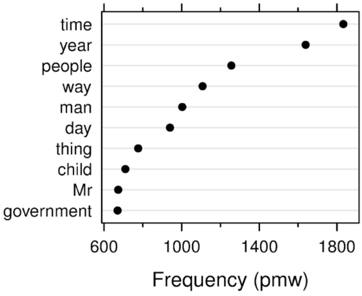


Figure 1. A simple dot plot showing the relative frequency of the 10 most frequent nouns in the BNC; data from Leech et al. (2001)

Simple dot plots can be extended. Figure 2 illustrates a number of additional features and defines the relevant terminology (largely borrowed from Cleveland 1994: 21–22). The data are from a study comparing British (BrE) and American English (AmE) newspaper texts regarding the preference for (orthographically) regular verb forms (e.g. *learned* vs. *learnt*) in simple past and past participle contexts (Levin 2009).

The *main panel* compares two groups (BrE vs. AmE) using different *plotting symbols*. These are *superposed* – that is, plotted on the same line – and labeled in the *key* at the top (arranged to match the major pattern in the plot). *Data labels* on the vertical scale list the verbs, which are ordered by relative frequency in BrE, increasing from bottom to top. The (optional) *appended panel* on the right expresses the *comparison* between BrE and AmE directly by plotting the differences. A *reference line* marks zero, which signals no difference, a relevant reference value. *Tick marks* point outward and are also drawn at the top to facilitate *table look-up*. Difference estimates are indicated by filled circles and include *error bars* showing 95% confidence intervals as a measure of statistical uncertainty. Error bars are explained in the *scale label*.

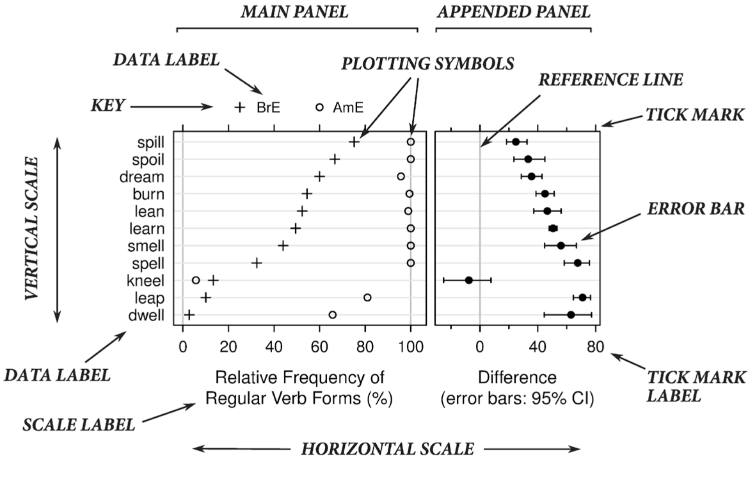


Figure 2. Elements of the dot plot: Terminology and style of presentation borrow heavily from Cleveland (1994); data from Levin (2009)

4. Advantages

The most common graphical display of labeled data is the bar chart, which has three variants: the simple, grouped, and stacked bar chart. It can be replaced by the dot plot in many of its established uses, which often produces a more effective display. This section discusses advantages of dot plots over bar charts.

Aesthetic minimalism

One of the principles of graphical design outlined by Tufte (2001) is the minimization of redundant visual information. Redundancy is expressed with the data-ink ratio, the ratio of "the non-erasable core of a graphic" to "the total ink used to print the graphic" (2001: 93). Using a single prominent symbol to show the data, the dot plot avoids superfluous visual elements. While there is no empirical evidence for the superiority of a high data-ink ratio (Spence 1990; Gillan & Richman 1994; Siegrist 1996), eliminating redundant ink yields a less cluttered graph and thus clear vision. Especially in multivariate displays this is an advantage over grouped or stacked bar charts. Figure 3 shows two variants of a grouped bar chart of Levin's (2009) results, both of which produce a more cluttered display compared to Figure 2.

Horizontal format

By convention, quantities are often plotted vertically. A horizontal orientation, however, yields four practical advantages: (i) the data labels are shown horizontally and are thus easy to read; (ii) long data labels do not require abbreviations or rotation (cf. Figure 3), which may slow down or even interfere with the decoding of information; (iii) the display can be extended comfortably to show a large number of data points (cf. Figure 9); (iv) the amount of (vertical) space needed for the graph can be reduced without affecting the resolution of the display. The horizontal format, however, yields a number of important limitations of this display method (see Section 7).

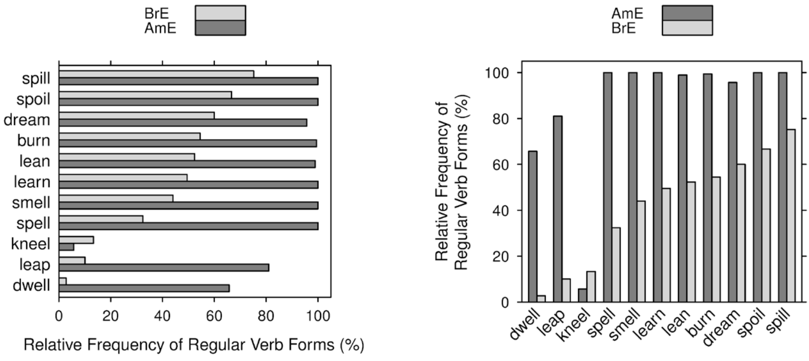


Figure 3. The data from the main panel in Figure 2 shown as a horizontal and vertical grouped bar chart

Resolution

Dot plots offer several benefits in terms of resolution. The issue of axis scaling – that is, whether scaling to zero is necessary – has received much attention in the literature. In science, there appears to be consensus that excluding zero from the scale is often desirable. "Zooming in" by rescaling can greatly enhance the resolution of a display and thus facilitate perception of the variation in the data (Tukey 1977: 51; Cleveland 1994: 92; Wainer 1997, 2009). On the other hand, it has also been argued that such rescaling is inherently misleading (e.g. Huff 1954; Krämer 2001). However, this partly depends on the type of display chosen (Robbins 2005). Bar charts use position (end of bar) as well as size (length and area) to encode numeric values. Without a baseline of measurement, the length of a bar encodes meaningless information and indeed provides misleading visual cues by exaggerating actual differences. Dot plots only use position; the distance to the left side of the graph is further de-emphasized by drawing light horizontal lines across the graph.

The resolution of a graph can be greatly affected by skew, where – due to a few outliers – most data points are crammed into a small part of the graph. Two remedies are data transformation and the use of a scale break. In contrast to conventional scale breaks (i.e. two short parallel lines intersecting the axis), a full scale break divides the graph into two panels, each with a full frame and its own scale (Cleveland 1984). The visually salient discontinuity arguably discourages pattern or continuity perception across the break. The more frequently applied strategy, however, is data transformation. Logarithms are a particularly useful tool when the data are skewed towards large values or when relative differences are of interest. When graphing on a logarithmic scale, dots should be used; the length of bars would provide meaningless information since a log scale has no logical baseline or origin. Figure 4 illustrates the use of a full scale break and a logarithmic transformation to the display of the 10 most frequent verbs in the BNC (Leech et al. 2001). Due to the dominance of the primary verbs (*be, do, have*), even the log transformation does not contribute much to our assessment of the variability among the lower-frequency verbs. In this case, the use of a full scale break helps.

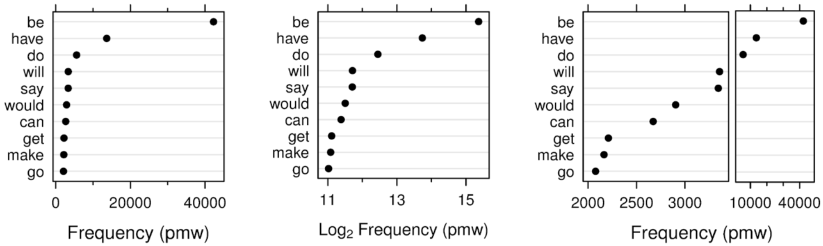


Figure 4. The 10 most frequent verbs in the BNC shown on the original scale, a $\log_2$ scale and using a full scale break; data from Leech et al. (2001)

Error bars

In many cases it is useful for point estimates to include information on statistical variation (Wilkinson & Task Force on Statistical Inference 1999). Such measures are typically indicated with error bars, which may denote different types of information (e.g. standard deviations, standard errors, a confidence interval or a percentile interval). An advantage of dot plots over bar charts is the fact that they produce a more effective presentation of error bars (Cleveland & McGill 1984; Schnell 1994; Wainer 2009). Figure 5 provides three displays of the difference scores that are shown in the appended panel of Figure 2. The principle of *discriminability* and the Gestalt law of *similarity* facilitate *detection* and *assembly* in the plot on the left. Point and interval estimates are visually discriminable, which makes it easy to focus on one type of estimate while mentally muting the other. In bar charts this is more difficult due to the similarity of geometric elements (right-angled linear segments with the same orientation). Minimization of ink adds *salience* to error bars and point estimates, which further facilitates *comparison* and assessment of the variability between verbs. The estimates for each verb are also more easily perceived as a single visual unit due to their point and axis symmetry (Gestalt law of *good form*).

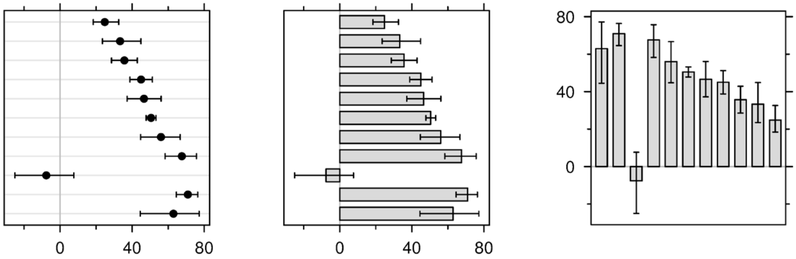


Figure 5. Error bars: Dot plot vs. horizontal and vertical bar chart

Interval scales

Quantitative variables are divided into interval- and ratio-scaled measures, depending on whether there is an absolute zero. While ratio variables only take on non-negative numbers, an interval scale allows positive and negative quantities (e.g. difference scores and correlation coefficients). Bar charts are ill-suited for interval scales, especially if positive and negative values occur in the same plot (cf. Figure 5). As pointed out by McElreath (2016: 203), the only information added by bars – at the expense of a more cluttered display – is "which way to zero". Moreover, error bars extending beyond zero yield an odd appearance (cf. Figure 5). The lengths of bars also encourage ratio comparisons ("A is about twice as large as B"), which may not be warranted on interval scales (e.g. in the case of correlation coefficients).

Pattern perception

When the plotted categories are grouped, the dot plot usually outperforms the divided and grouped bar chart. A comparison of Figures 2 and 3 shows that grouped bar charts quickly become cluttered, which interferes with *pattern perception* (Robbins 2005). In dot plots, successful superposition facilitates Gestalt-like perception of the groups – that is ,they can be visually assembled while mentally filtering out the other elements (Cleveland 1994). Bertin (1983: 67) calls this “selective perception”, noting that “[t]he eye must be able to isolate all the elements of [a] category, disregard all the other signs, and perceive the image formed by the given category”.

Estimation

Experimental research into graphical perception has identified a number of elementary perceptual tasks that are used to visually extract quantitative information from a display. Examples of such tasks are position, length, angle and area judgments. The visual decoding of dot plots involves position judgments along a common scale. This elementary perceptual task produces more accurate *estimation* than length or angle judgments, which are used in decoding bar charts and pie charts, respectively (Cleveland & McGill 1984). However, performance differences between position, length and angle judgments may be relatively small (Carswell 1992).

In sum, several arguments suggest that the widely used bar chart can in many cases be constructively replaced by a dot plot.

5. Design

There are different options for the design and extension of dot plots. This section illustrates a number of add-ons and discusses construction principles aiming to optimize the resulting display. While such fine-tuning is particularly important for presentation graphs, most of the following considerations are also relevant for the use of dot plots in data analysis.

Order

Location in the two-dimensional space is a powerful visual cue and can be attended to easily and selectively (Kubovy 1981). Dot plots should thus make use of the vertical dimension by (re-)ordering categories or groups in specific ways. If the categories have no logical arrangement, data-based ordering (often according to value or size) facilitates information processing and reveals additional structure in the data (Bertin 1983; Schmid 1983; Wainer 1997). This also applies to multipanel displays and the use of *superposition,* where different options for ordering exist. Ordered symbols are more easily perceived as belonging to the same group (Gestalt law of *proximity* and *good form*). The data analyst should try different arrangements to foreground different gestalts and comparisons, which is likely to uncover different aspects and patterns in the data.

Multipanel conditioning and superposition

Additional categorical variables can be incorporated into dot plots by means of superposition or juxtaposition. In essence, these are different plotting strategies for the comparison of (sub-)groups. While superposed groups are shown in the same panel (cf. Figure 2), juxtaposition involves the use of multiple panels to plot subsets of the data (cf. Figure 11). In general, multipanel conditioning is a powerful method for the display of multivariate data sets (Becker et al. 1996). It is important to note that the two strategies are complementary approaches to the display of multivariate data sets. In general, however, superposition facilitates comparisons *between* groups; juxtaposition, on the other hand, strives for clear vision and allows for better comparison *within* groups. When the number of groups is small, superposition may be more effective than multipanel conditioning (Sarkar 2008).

Plotting symbols

The choice of plotting symbols should allow for good visual detection and assembly. In a simple dot plot, filled circles (●) are recommended since they are salient and combine well with error bars. If two groups are compared in the same panel, the choice of plotting symbols depends on whether overplotting occurs. When there is no overplotting, filled and empty circles (● ○) are a good choice. In the case of overplotting, empty circles and crosses (○ +) allow for better pattern perception (Cleveland 1994). Their distinct pre-attentive attributes ensure excellent texture discrimination (Malik & Perona 1990). Ease of detection and assembly allow the viewer to focus on one group while backgrounding the other. Moreover, salient filled circles (●) can then be used in appended panels to directly show key comparisons (cf. Figure 2). Empty circles and crosses (○ +) may also serve as iconic symbols, signaling presence/absence of a certain attribute (cf. Figure 8). In addition, other symbols may make sense. Letters, for instance, make it easier to remember the groups or categories shown, saving time that would otherwise have been spent looking back and forth between key and data (cf. Figure 8). However, the set of symbols should still be sufficiently discriminable (on the discriminability of graphemes see Lewandowsky & Spence 1989).

Appended panels

A particularly useful add-on for dot plots are appended panels (cf. Amit et al. 2008; see also Heiberger & Holland 2015: 566). Despite its superficial similarity, this plotting strategy is conceptually different from multi-panel conditioning. Appended panels do not display a different subset of the data, but rather add more information on the data set plotted in the main panel. While there are many possible uses for appended panels, they seem particularly valuable for directly showing focused comparisons between two groups in the main panel. Such comparisons can be expressed using various types of effect sizes such as difference or ratio measures. Alignment along a common scale makes it much easier to compare effects across categories on the y-axis (e.g. the different verbs in Figure 2). Since different effect size measures may offer different perspectives on the same comparison, it may make sense to append more than one panel (cf. Figure 8). Inferential information can be added to effect size estimates in the form of confidence intervals, which indicate the degree of uncertainty associated with the estimates shown (see Figures 2, 9 and 10).

Error bars

Several options exist for the design of error bars, differing in the way interval limits are marked and as to how many intervals are shown for each point estimate. Figure 6 shows several variants. The most widely used type of error bar is single-tiered with the upper and lower limit marked by crossbars. The use of crossbars has met with criticism since it draws attention to the endpoints of the interval. At any rate it appears reasonable to limit crossbar length to the diameter of the plotting symbol of the point estimates. Error bars may also display several intervals for the same estimate. Cleveland (1994), for instance, suggested two-tiered error bars for showing different confidence levels. Outer tiers may also be used to add interval limits that are adjusted for multiple comparisons (Tukey 1993; cf. Figure 9). As illustrated in Figure 6, inner intervals can be delimited using crossbars (cf. Cleveland 1985: 226), line width (cf. Gelman & Hill 2007: 497) or shading (cf. Harrell 2015: 282). While this appears to be a matter of taste, the use of different line types (more specifically, solid and dashed lines) should be avoided as dash patterns may lead to minor inaccuracies in the boundary locations of the inner and outer tiers (see Kastellec & Leoni 2007: 759 for an example).

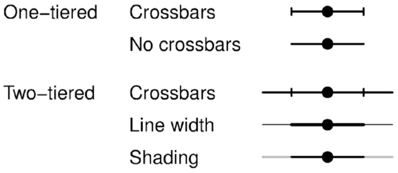


Figure 6. Design options for one-tiered and two-tiered error bars

Dodging

When adding error bars to panels with superposed plotting symbols, overlap is an issue. A simple strategy is to use vertical displacement (see middle panel of Figure 11). In Wickham's (2013) *ggplot2* package for R, this strategy is called "dodging". Plotting symbols and error bars are relocated above and below the light horizontal line, which avoids error bar overlap while still preserving assembly and pattern perception.

Logarithms

Logarithms are a very useful tool for data analysis and visualization. Plotting on a log scale shows relative rather than absolute differences. While logarithms can be expressed using different log bases, it is important to note that the choice of base does not affect the *physical* information in the plot: the same pattern occurs regardless of whether log base 2, *e* or 10 is used (these are typical choices). What changes is the *scale* information, that is, the tick mark labels. The base should be chosen to facilitate *table look-up*. This includes recovering the original values of the points plotted and, more importantly, making comparisons, that is, *estimating* the relative difference between two points plotted. The viewer can be assisted in making these judgments by adding original units to the tick marks at the top (cf. Figure 9).

Lines and color

If more than two groups are superposed in the same panel, it becomes increasingly difficult for the viewer to *detect* and *assemble* groups. *Pattern perception* may then be facilitated by using color or linking the points with lines (Gestalt law of *connectedness*). With the addition of lines, the graphical display approaches the fuzzy category boundary to parallel coordinate plots (see Unwin 2015: 99–130) and line plots (sometimes called interaction plots). The use of lines is further discussed in Section 7.

6. Applications

This section will illustrate the application of dot plots to different types of descriptive and inferential analyses of linguistic data, demonstrating most of the design options discussed above.

Meta-analysis

The term meta-analysis refers to a set of techniques for combining evidence from different studies on the same or similar issues (Cumming 2012). Graphs play an important role in meta-analysis. A frequently employed display type is the forest plot (Lewis & Clarke 2001), which allows the researcher to visually assess effect estimates and confidence intervals reported in the literature. It thus provides a graphical synthesis of the empirical evidence available (see Borenstein et al. 2009 for many examples). Proper meta-analyses also condense the evidence into a single effect size estimate with a (usually much narrower) confidence interval. A simple visual summary, however, is a useful starting point since it allows for a contextualization of new findings, yielding a more solid basis for their interpretation (recall Tufte's quote on quantitative reasoning; Section 2). Forest plots are in fact very similar to dot plots but include a few additional features such as the variation of the size of plotting symbols to signal the degree of uncertainty associated with a particular point estimate (see Lewis & Clarke 2001; Cumming 2012).

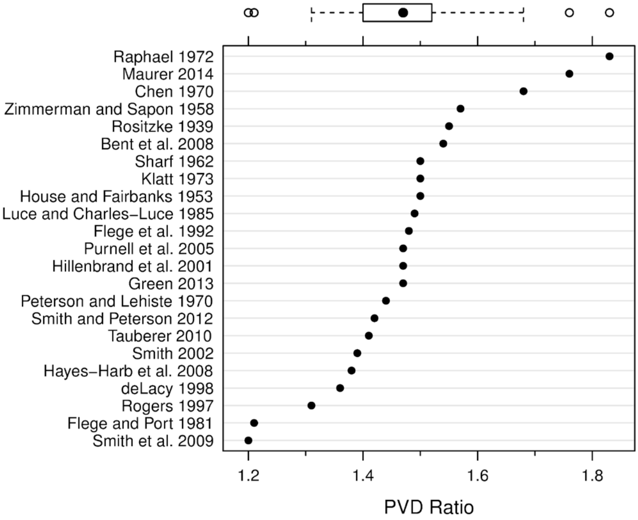


Figure 7. A visual summary and comparison of the results from different studies on the same phenomenon (PVD effect in American English)

Figure 7 summarizes empirical estimates of the PVD (preceding vowel duration) effect in American English. Minimal pairs differing in final obstruent voicing (e.g. *bad-bat, peas-peace*) are primarily distinguished by the duration of the preceding vowel, which is longer before voiced obstruents (*bad*, *peas*). This PVD effect can be expressed as a duration ratio. Figure 7 shows the estimates obtained in 23 studies in increasing order from bottom to top. A boxplot has been added to show the distribution of the values. The literature is in fairly good agreement that the ratio ranges somewhere between 1.4 and 1.5. A few studies have reported particularly high or low values, which would prompt us to study their methods section in more detail to identify possible confounding variables.

Figure 8 shows another application of the dot plot to a simple research synthesis. It gives an overview of the empirical evidence on the voice onset time (VOT) of voiced and voiceless stops in American English. VOT, the duration of the interval between stop release and onset of voicing, is an acoustic correlate of the voicing distinction in initial stops. The main panel shows the measurements reported in each study, ordered by the overall average VOT, increasing from bottom to top. Letters (more precisely: IPA symbols) serve as plotting symbols, which facilitates *table look-up*. A reference line is included at zero, an important reference value for these data. The box plots above the main panel compare the distribution of voiced and voiceless consonants. While there is little variation across studies regarding voiced stops, VOT measurements for voiceless stops differ drastically. Further, it is obvious that VOT varies systematically with place of articulation: velar stops /k,g/ show the largest, bilabial stops /p,b/ the smallest values.

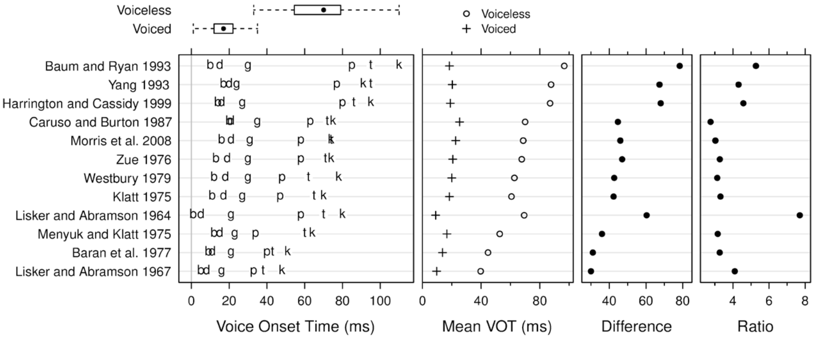


Figure 8. A visual summary of the results of different studies on the same phenomenon (VOT in American English stops)

Three appended panels provide different perspectives on the data in the main panel. The second panel shows the average VOT for voiced and voiceless stops. The plotting symbols indicate the presence (+) or absence (○) of voicing (i.e. the feature [±voice]). The difference in variability between the two categories is even more apparent in this display. To directly express the effect of [±voice] on VOT, difference or ratio scores may be used. These facilitate the comparison across studies and are shown in the appended panels further on the right. To increase resolution, the x-axis does not include zero in these two rightmost panels. It is clear that difference measures covary with overall average VOT, which in turn mostly reflects the VOT in voiceless stops. Ratio measures appear to somewhat control for this effect and may thus be the preferred measure for comparing results across studies. The panels on the right also force the viewer to note that one study clearly sticks out – a finding in need of an explanation.

Corpus data analysis

There are two types of data that frequently arise in corpus linguistics: binary and frequency outcomes. While frequency data reflect the number of occurrences of an event (e.g. word) during a period of observation (e.g. a text or a corpus), binary data stem from variables comprising two categories (or levels), such as regular vs. irregular verb form. Characteristically, corpus-based studies involve two types of comparisons. Commonly, researchers contrast (sub-)corpora representing different varieties of language (such as spoken vs. written) or populations of speakers (e.g. native vs. non-native). On the other hand, it is also typical to investigate several items (lexemes or constructions of any kind). We may therefore distinguish between the comparison of groups and items.

Figure 9 shows an application of the dot plot to the analysis of corpus frequencies (counts). The data are from Granger & Paquot (2008), a study on verb usage in learner and expert academic writing. Counts from two corpora representing non-native and native speaker academic writing were compared. The plot shows the “top 50 underused” verbs in learner academic writing, which were selected based on the likelihood ratio test statistic. There are two types of comparisons: between groups (learners vs. experts) and items (verbs).

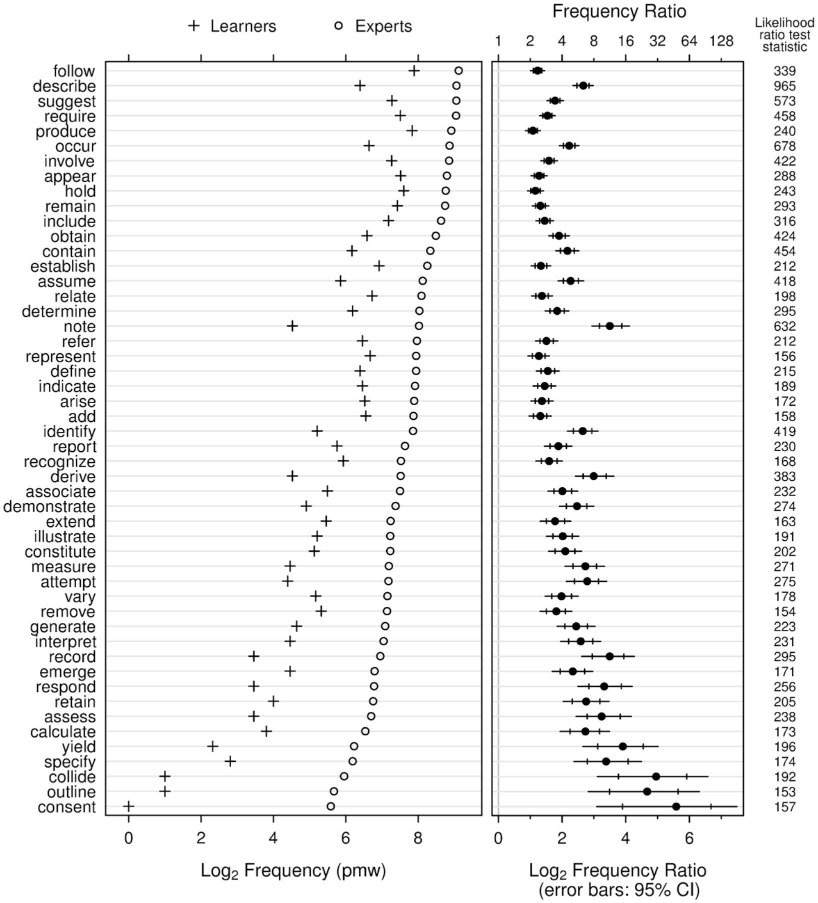


Figure 9. Corpus frequencies: Underrepresented verbs in learner academic writing; data from Granger & Paquot (2008). Inner error bars show individual 95% CIs; outer error bars show 95% CIs adjusted for multiple comparisons

The main panel shows the relative frequency estimates (per million words) for the verbs, which are ordered by their frequency in expert academic writing. Frequency is shown on a log2 scale, which ranges from 0 (2^0 = 1 pmw) to about 9 (2^9 = 512 pmw). The appended panel shows the log2 frequency ratio (more precisely: the logarithm (to base 2) of the ratio of the absolute frequencies), which expresses the degree of underuse in learner writing. These ratio scores are shown on a logarithmic scale, and also translated to the original ratio measures by adding the respective tick mark labels at the top. Most verbs are around 2 to 8 times more likely to occur in native speaker expert writing. A handful of verbs are severely underrepresented in learner writing (e.g. *collide, outline, consent*). Information on statistical uncertainty is added in the form of two-tiered error bars. While the inner tiers (delimited by crossbars) show individual 95% confidence intervals, the outer tiers show 95% CIs adjusted for multiple comparisons using the Bonferroni procedure (i.e., showing 1 – α/50 = 99.9% CIs). The likelihood ratio test statistics, which are added at the right margin of the plot, show that comparisons based on such measures may miss important information in the data (such as the underrepresentation of *collide, outline* and *consent*, which have relatively low test statistics due their sparse occurrence, especially in learner writing).

Figure 10 shows an application of the dot plot to the analysis of binary outcomes. The data are from a study by Mondorf (2009) on the variation between synthetic and analytic comparative forms of adjectives in British and American English newspaper writing. Adjectives may form the comparative synthetically with an inflectional suffix (*prouder, purer*) or analytically with *more* (*more proud, more pure*). Mondorf (2009) investigated differences between the varieties in the preference for a particular formation strategy in a number of monosyllabic adjectives. This study thus also involves two types of comparison: between groups (BrE vs. AmE) and items (adjectives).

The results for 15 adjectives are shown in Figure 10. The main panel plots the percentage of analytic comparatives; items are ordered by their relative frequency in AmE, increasing from bottom to top. The appended panel shows the difference in relative frequency between AmE and BrE. There appears to be a bipartition into adjectives with predominantly analytic comparatives (at the top) and those preferring a synthetic form (towards the bottom). Except for *free* and *true,* AmE always shows a stronger tendency towards analytic comparatives. The absolute difference in relative frequency typically ranges from 0 to 20%, with *sour* being a notable outlier (a difference of almost 60% in absolute terms). Reference lines mark the limits of 0 and 100 in the main panel and the reference value of 0 in the appended panel, which denotes equal distribution of synthetic/analytic forms in the two varieties.

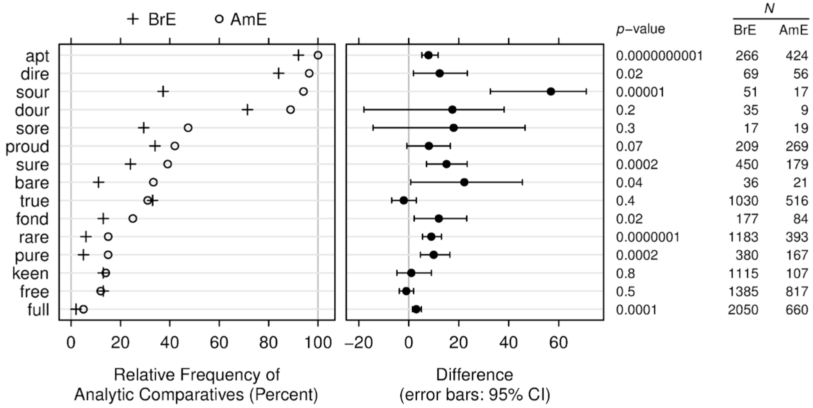


Figure 10. Analysis of a binary outcome: Synthetic vs. analytic comparatives in British vs. American English; data from Mondorf (2009)

P-values from corresponding likelihood ratio tests are added at the right margin; they are in good agreement with the 95% confidence intervals shown in the appended panel. The *p*-values have been rounded up to one significant digit, which produces a semi-graphic representation similar to a stem-and-leaf display (Tukey 1977). Note the unusually large difference between British and American English for *sour*. This extreme divergence is not directly reflected in its *p*-value since this type of measure conflates effect magnitude and sample size. The absolute counts for each adjective are shown to the right of the plot. These correlate with the widths of the confidence interval. Clearly, the corpus contained relatively few tokens of *sourer/more sour*. Thus, while *p*-values collapse effect and sample size into a single measure, effect sizes with confidence intervals allow the researchers to compare and interpret both measures, effect magnitude and statistical uncertainty. This is a strong argument for the preference of confidence intervals over p-values (see also Gardner & Altman 2000).

Application in statistical modeling

Graphical methods play an important role in statistical modeling. Especially in multivariate models, it is difficult for the analyst (as well as the audience) to make sense of tables of coefficients, which are the default output of most statistical software. Indeed, as noted by McElreath, statistical models have "terrible people skills" (2016: 232). Among many other graph types, dot plots have emerged as a particularly suitable aid in model understanding and comparison. As such, model-derived quantities that are translated into graphical form include regression coefficients, test statistics and information criteria. Rather than discuss particular examples of dot plots in statistical modeling, this section will hint at a range of applications and include textbook references for further study.

One strategy is to plot regression coefficients with their associated measures of statistical uncertainty (see Kastellec & Leoni 2007: 765; McElreath 2016: 375, 401). This strategy foregrounds the effect of the predictor and prompts the viewer to compare coefficients rather than *p*-values (or asterisks). If input variables differ in level of measurement and scale, this raises the issue of comparability. For least squares regression, corrective actions include the standardization of regression coefficients (Fox 2016: 100–102; see also Gelman 2008). In logistic regression models, dot plots can be applied for the comparison of predictors based on regression coefficients (Gelman & Hill 2007: 306), odds ratios (Harrell 2015: 282), test statistics (Harrell 2015: 280), and average predictive comparisons (Gelman & Hill 2007: 466–473).

Further comparisons in statistical modeling involve quantities derived from different models for the same data. Such differences may arise from the number of predictors included (Gelman et al. 2013: 423; McElreath 2016: 202) or the fitting of mathematically and/or conceptually different models (Gelman & Hill 2007: 202, 473; Gelman et al. 2013: 400). Further, graphical displays may be used to compare subgroups (Gelman & Hill 2007: 338) or serve as an aid to model comparison using information criteria (McElreath 2016: 199).

7. Limitations

Like other graph types, dot plots have limitations that need to be considered when choosing between different chart types.

Unfamiliarity

One obstacle the dot plot faces is its unfamiliarity to most viewers, which may violate the principle of *appropriate knowledge*. Recognition of the graph type is a critical step in the processing and comprehension of a graphical display. As Kosslyn (1985: 507) notes, "if one has never seen a display type before, it is a problem to be solved – not a display to be read". While the use of dots to encode numerical values in simple displays should pose no problems, more elaborate constructions including superposition and multipanel conditioning may be more demanding for certain audiences. The use of dot plots thus requires reflection on the "graphicacy" (Keen 2010) of the intended audience as well as the time available for graph comprehension (principle of *capacity limitations*). The limitation of unfamiliarity, however, does not apply to the application of dot plots in data analysis.

Cognitive fit

A limitation that applies to the application of the dot plot in data analysis and presentation is the issue of cognitive fit between graph and data (Vessey 1991): the type of display chosen should be compatible with the type of information shown (principle of *compatibility*). Since dot plots show categories on the vertical axis, they are ill-suited for depicting independent variables that are by convention shown on a horizontal axis. Examples are time series and quasi-time differences, such as time trends or variables reflecting age groups, developmental stages or pre- and post-test scores. Figure 11 shows results from an experimental study on plural overregularization in English children's production of irregular plural nouns, for instance **mouses* instead of *mice* (Ramscar et al. 2013). The researchers hypothesized that training on regular plurals would lead to an increase in overregularization in younger children, but to a decrease in older children. The degree of overregularization was measured before (pretest) and after training (posttest) on a scale from -1 (no overregularization) to 1 (overregularization). The bar chart in Figure 11 (left panel) resembles the graph used by the authors to display their results. Due to the interval scale, a bar chart is less suitable (the origin at -0.6 is arbitrary). The dot plot in the middle is a first attempt at producing a more satisfactory display (Sönning 2014) but fails to clearly communicate the experimental results. Since these data involve change over time (as a result of training), the two time points (pretest and posttest) should be shown on the horizontal axis (principle of *compatibility*), which could be conveniently achieved by a line plot, for instance (see right panel of Figure 11). Line plots have a further advantage over other chart types: *assembly* can be greatly assisted by the use of lines (Gestalt law of *connectedness*) and *table look-up* is facilitated by direct labeling of these lines (Gestalt law of *proximity*). As a result, there is no need for a key, which accelerates the decoding of information (Milroy & Poulton 1978; Parker 1983, cited in Pinker 1990: 114).

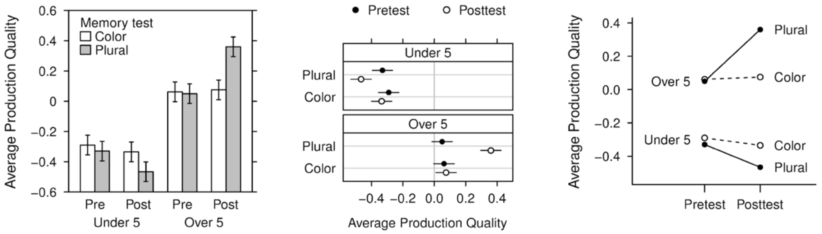


Figure 11. Graph types and cognitive fit: Pretest and posttest scores in two age groups (under 5 vs. over 5) and two tests (color vs. regular plural) shown as a grouped bar chart, a dot plot and a line plot; data from Ramscar et al. (2013)

8. Software

The plots in this paper were drawn in *R* (R Core Development Team 2016) using the packages *lattice* (Sarkar 2008) and *latticeExtra* (Sarkar & Andrews 2016). There is a short tutorial on the construction of dot plots using lattice in the online appendix (www.bit.ly/malt-dotplot-lattice). This appendix also includes dot plot templates for Microsoft Excel, which enable the user to easily construct dot plots (including the use of superposition and appended panels) by copy-and-pasting their data into spreadsheet templates (www.bit.ly/malt-dotplot-excel). Of course, using a template means that there are fewer design options than in *R*. A short instruction manual is also provided online (www.bit.ly/excel-dotplot-instructions).

9. Conclusion

In this contribution, I have argued that the dot plot is a flexible tool for visualizing different types of numeric values with descriptive labels: raw data, frequencies, descriptive measures and model parameters. It is able to replace the bar chart in most of its established uses and likely to produce a more effective display of the data. This paper has demonstrated advantages of the dot plot, illustrated principles for its design and extension to multivariate data sets and exemplified their application to quantitative data in linguistics. Dot plots are a useful tool for data visualization. They should be used more often.

References

Amit, Ohad, Richard M. Heiberger & Peter W. Lane. 2007. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics* 7. 20–35.

Becker, Richard A., William J. Cleveland & Ming-Jen Shyu. 1996. The visual design and control of trellis displays. *Journal of Computational and Graphical Statistics* 5. 123–155.

Bertin, Jacques. 1983. *Semiology of graphics*. Madison: University of Wisconsin Press.

Borenstein, Michael, Larry V. Hedges, Julian P.T. Higgins & Hannah R. Rothstein. 2009. *Introduction to meta-analysis*. Chichester: Wiley.

Carswell, C. Melody. 1992. Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Human Factors* 34(5). 535–554.

Cleveland, William S. 1984. Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician* 38(4). 270–280.

Cleveland, William S. 1993. A model for studying display methods of statistical graphics. *Journal of Computational and Statistical Graphics* 3. 323–343.

Cleveland, William S. 1994. *The elements of graphing data*. Summit: Hobart Press.

Cleveland, William S. & Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association* 79. 531–554.

Cumming, Geoff. 2012. *Understanding the new statistics: Effect sizes, confidence intervals and meta-analysis*. New York: Routledge.

Fienberg, Stephen E. 1979. Graphical methods in statistics. *American Statistician* 33(4). 165–178.

Fox, John. 2016. *Applied regression analysis and generalized linear models*. Thousand Oaks: Sage.

Gardner, Martin J. & Douglas G. Altman. 2000. Confidence intervals rather than P values. In Douglas G. Altman, David Machin, Trevor N. Bryant & Martin J. Gardner (eds.), *Statistics with confidence*, 15–27. London: BMJ Books.

Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27. 2865–2873.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical linear models*. Cambridge: Cambridge University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin. 2013. *Bayesian data analysis*. Boca Raton: CRC Press.

Gillan, Douglas J. & Edward H. Richman. 1994. Minimalism and the syntax of graphs. *Human Factors* 36(4). 619–644.

Granger, Sylviane & Magali Paquot. 2008. Lexical verbs in academic discourse: A corpus-driven study of learner use. In Maggie Charles, Diane Pecorani & Susan Hunston (eds.), *Academic writing: At the interface of corpus and discourse*, 193–214. New York: Continuum.

Harrell, Frank E. Jr. 2015. *Regression modeling strategies*. New York: Springer.

Heiberger, Richard M. & Burt Holland. 2015. *Statistical analysis and data display: An intermediate course with examples in R*. New York: Springer.

Huff, Darrell. 1954. *How to lie with statistics*. New York: W. W. Norton.

Jacoby, William G. 2006. The dot plot: A graphical display for labeled quantitative values. *The Political Methodologist* 14(1). 6–14.

Kastellec, Jonathan P. & Eduardo L. Leoni. 2007. Using graphs instead of tables in political science. *Perspectives on Politics* 5. 755–771.

Keen, Kevin J. 2010. *Graphics for statistics and data analysis with R*. Boca Raton: CRC Press.

Kosslyn, Stephen M. 1985. Graphics and human information processing: A review of five books. *Journal of the American Statistical Association* 80. 499–512.

Kosslyn, Stephen M. 2006. *Graph design for the eye and mind*. Oxford: Oxford University Press.

Krämer, Walter. 2001. *So lügt man mit Statistik*. München: Piper.

Krug, Manfred, Ole Schützler & Valentin Werner. 2016. Patterns of linguistic globalization: Integrating typological profiles and questionnaire data. In Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja & Sarah Chevalier (eds.), *New Approaches to English Linguistics: Building Bridges,* 35–66. Amsterdam: Benjamins.

Kubovy, Michael. 1981. Concurrent pitch segregation and the theory of indispensable attributes. In Michael Kubovy & James R. Pomerantz (eds.), *Perceptual organization*, 55–98. Hillsdale: Erlbaum.

Leech, Geoffrey, Paul Rayson & Andrew Wilson. 2001. *Word frequencies in written and spoken English*. London: Longman.

Levin, Magnus. 2009. The formation of the preterite and the past participle. In Günter Rohdenburg & Julia Schlüter (eds.), *One language, two grammars? Differences between British and American English,* 60–85. Cambridge: Cambridge University Press.

Lewandowsky, Stephan & Ian Spence. 1989. Discriminating strata in scatterplots. *Journal of the American Statistical Association* 84. 682–688.

Lewis, Steff & Mike Clarke. 2001. Forest plots: Trying to see the wood and the trees. *British Medical Journal* 322. 1479–1480.

Malik, Jitendra & Pietro Perona. 1990. Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A* 7. 923–932.

McElreath, Richard. 2016. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton: CRC Press.

Milroy, Robert & E. Christopher Poulton. 1978. Labelling graphs for improved reading speed. *Ergonomics* 21. 55–61.

Mondorf, Britta. 2009. Synthetic and analytic comparatives. In Günter Rohdenburg & Julia Schlüter (eds.), *One language, two grammars? Differences between British and American English,* 86–107. Cambridge: Cambridge University Press.

Palmer, Stephen E. & Irvin Rock. 1994. Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin and Review* 1. 29–55.

Pinker, Steven. 1990. A theory of graph comprehension. In Roy Freedle (ed.), *Artificial intelligence and the future of testing*, 73–126. Hillsdale: Erlbaum.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ramscar, Michael, Melody Dye & Stewart M. McCauley. 2013. Error and expectation in language learning: The curious absence of mouses in adult speech. *Language* 89(4). 760–793.

Robbins, Naomi B. 2005. *Creating more effective graphs*. Hoboken: Wiley.

Sarkar, Deepayan. 2008. *Lattice: Multivariate data visualization with R*. New York: Springer.

Sarkar, Deepayan & Felix Andrews. 2016. *latticeExtra: Extra Graphical Utilities Based on Lattice*. R package version 0.6-28. https://cran.r-project.org/web/packages/latticeExtra/latticeExtra.pdf (30 November, 2016.)

Schmid, Calvin F. 1983. *Statistical graphics: Design principles and practices*. New York: Wiley.

Schnell, Rainer. 1994. *Graphisch gestützte Datenanalyse*. München: Oldenbourg.

Schützler, Ole. Forthcoming. A corpus-based study of concessive conjunctions in three L1-varieties of English. In Isabelle Buchstaller & Beat Siebenhaar (eds.), *Language variation – European Perspectives VI: Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8)*, Leipzig, May 2015. Amsterdam: Benjamins.

Siegrist, Michael. 1996. The use or misuse of three-dimensional graphs to represent lower-dimensional data. *Behaviour and Information Technology* 15. 96–100.

Sönning, Lukas. 2014. The dot plot: A fine tool for data visualization. Paper presented at *Advances in Visual Methods for Linguistics* (AVML). University of Tübingen.

Spence, Ian. 1990. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance* 16. 683–692.

Tufte, Edward R. 1990. *Envisioning information*. Cheshire: Graphics Press.

Tufte, Edward R. 2001. *The visual display of quantitative information*. Cheshire: Graphics Press.

Tukey, John W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, John W. 1993. Graphic comparisons of several linked aspects: Alternatives and suggested principles. *Journal of Computational and Graphical Statistics* 2. 1–33.

Unwin, Antony. 2015. *Graphical data analysis with R*. Boca Raton: CRC Press.

Vessey, Iris. 1991. Cognitive fit: A theory-based analysis of the graphs vs. tables literature. *Decision Sciences* 22. 219–241.

Wainer, Howard. 1997. *Visual revelations*. Mahwah: Erlbaum.

Wainer, Howard. 2009. *Picturing the uncertain world*. Princeton: Princeton University Press.

Ware, Colin. 2013. *Information visualization: Perception for design*. Amsterdam: Elsevier.

Werner, Valentin & Robert Fuchs. 2016. The present perfect in Nigerian English. *English Language and Linguistics* First View.

Wertheimer, Max. 1938. Laws of organization in perceptual forms. In Willis D. Ellis (ed.), *A source book of Gestalt psychology*, 71–88. London: Routledge & Kegan Paul.

Wickham, Hadley. 2013. *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wilkinson, Leland & Task Force on Statistical Inference. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* 54. 594–604.

Part 2

Bringing methods and linguistic theories together

A corpus-based quantitative approach to the study of morphological productivity in diachrony: The case of *samo*-compounds in Russian

Chiara Naccarato
University of Pavia and University of Bergamo

The present paper aims at investigating the productivity of the prefixoid *samo-* ('self') in Russian compounds from a diachronic perspective. In order to verify the hypothesis that the productivity of this prefixoid has grown over time, I consider the occurrences of *samo*-compounds in the *Russian National Corpus*, dividing the main corpus into four subcorpora, each one representing a particular time span: the 18th century, the 19th century, the 20th century and the period that lasts from the beginning of the 21st century to the present day. The approach chosen is quantitative in nature, and is based on the measure of "potential productivity" (Baayen & Lieber 1991; Baayen 1992, 1993), which is calculated by dividing the number of *hapax legomena* with a certain affix by the number of tokens with that affix. This measure, however, seems inadequate for the comparison of differently-sized corpora. To overcome this problem, I resort to parametric statistical models of frequency distribution known as LNRE (Large Number of Rare Events) models (Baayen 2001). These models, which allow extrapolating the expected values of types and *hapax legomena* with a given affix for arbitrary values of tokens, are implemented in the package zipfR (Baroni & Evert 2014), a tool for lexical statistics in R, which is used for this study.

"*Očen' mne nravitsja ètot russkij variant slovoobrazovanija pri pomošči «samo» — samo-let, samovar, samodejatel'nyj: vrode kak vse samo soboj bez našich usilij polučaetsja, samo!*"
"I really like that Russian word-formation pattern with «*samo*» — *samolet* (aircraft), *samovar* (samovar), *samodejatel'nyj* (amateur): as if everything happens without our efforts, by itself!" --- [Konstantin Serafimov, Èkspedicija vo mrak (1978–1996) // 1994]

1. Introduction

The paper studies the productivity of the Russian prefixoid[1] *samo-* ('self') and its change over time. The analysis is based on a large corpus of writ-

[1] A prefixoid is a word-initial affixoid. Affixoids are words that "have become similar to affixes in having a specialized meaning when embedded in compounds" (Booij 2010: 57). The constituent *samo-* can be described as a prefixoid because, despite being an unbound form, its behavior in compounds makes it similar to prefixes.

ten and spoken Russian, the *Russian National Corpus* (henceforth RNC, http://ruscorpora.ru/), and employs quantitative statistical methods to measure productivity.

This study contributes to showing that quantitative methods can help determine the fuzzy notion of morphological productivity, following the idea already suggested by many scholars that qualitative methods are not sufficient when investigating productivity: "(M)any researchers have abandoned the idea of a qualitative notion of productivity and have turned to the exact determination of [...] 'profitability'" (Plag 2006: 540), that is, productivity in a quantitative sense.

The paper is organized as follows: In Section 2, I present an overview of the quantitative approach to morphological productivity. In Section 3, I describe the data used for the study and the corpus pre-processing, that is, the selection of adequate items for the analysis. In Section 4, I show the quantitative analysis of the data and its results. In Section 5, I address the question of how productivity and lexicalization can coexist within one word-formation pattern. Finally, in Section 6, I discuss and summarize the findings.

2. A quantitative approach to morphological productivity

Many areas of linguistic research are increasingly oriented towards the use of quantitative empirical methods, which are sometimes essential to provide a satisfying analysis of certain linguistic phenomena. This is the case as regards studies concerning morphological productivity, which cannot dispense with the use of quantitative methods.

The distinction between qualitative and quantitative notions of productivity is discussed in Bauer (2001: 49, 205), where the author (following Corbin 1987) distinguishes the concepts of "availability" and "profitability" (*disponibilité* and *rentabilité* in Corbin's terms). In a qualitative approach, productivity is intended as a yes/no question: a word-formation process is either available or not for the creation of new words. In a quantitative approach, morphological productivity, that is, profitability, is understood instead as "the extent to which a morphological process may be employed to create new pertinent forms" (Plag 2006:

539). This notion of productivity is more focused on the actual use of a certain morphological process in performance and aims at determining the exact value corresponding to the productivity of the morphological process considered during a certain time span. However, as Plag (1999: 22) observes,

> quantitative and qualitative notions of productivity [...] are closely related. Thus the idea of potentiality, which is central to qualitative definitions of productivity, can be expressed in the statistical terms of probability.

The most widespread corpus-based quantitative approach to morphological productivity is represented by the works of Baayen and his collaborators (Baayen & Lieber 1991; Baayen 1992, 1993, 2001), who have elaborated on a number of corpus-based statistical measures of productivity. One of these measures is type-frequency V: the number of types with a given affix in a corpus of N tokens. As Bauer (2001: 144) suggests, this measure cannot provide information about the availability of a given word-formation process. It can only indicate that the given affix has been productive in the past. A second measure is “potential productivity” or “productivity in the narrow sense”, which is calculated by dividing the number of *hapax legomena* (V1) with a given affix by the number of tokens (N) with that affix: P = V1/N. According to Baayen & Lieber (1991: 809–810), this measure “estimates the probability of coming across new, unobserved types, given that the size of the sample of relevant observed types equals N”. This measure of productivity has been subject to criticism, since the value of P is highly dependent on the sample size (P is a function of N). Thus, when comparing differently-sized corpora, the measure P seems to be inadequate (Lüdeling et al. 2000; Evert & Lüdeling 2001; Gaeta & Ricca 2006; Štichauer 2009; Efthymiou et al. 2012). However, this measure can be applied if its value is calculated at equal token numbers, as I show in Section 4. A third measure proposed by Baayen & Lieber (1991: 819) and Baayen (1992: 124, 1993: 190) is “global productivity” (P*), which is a function of P and V. According to this measure, which is represented through a two-dimensional graph with the degree of productivity P on the horizontal axis and the value V on the vertical axis, productive affixes will have large values for P and V, while unproductive affixes will have low values for P and V. However, as pointed out both by Baayen himself (1992: 124) and Bauer (2001: 154), it is difficult to rank different affixes in terms of global productivity due to their disparate positions on the graph. Finally, another measure proposed by Baayen (1993: 192) is the "*hapax*-conditioned degree of productivity" (P*), which is calculated by dividing the number of *hapax legomena* with a given affix by the total number of *hapax legomena* in the corpus. However, as Bauer (2001: 155) points out,

> this measure asks 'What proportion of new coinages use affix A?' rather than asking 'What proportion of words using affix A are new coinages?' It is this latter which seems a more relevant question to ask.

Considering that productivity is about the creation of new words, the correlation between *hapax legomena* and productivity has often been criticized in view of the fact that *hapax legomena* are not necessarily new formations. However, as Gaeta & Ricca (2015: 847) suggest, "for corpora of many tenths of millions tokens, most hapaxes indeed turn out to be un-established words", and if the corpus data are manually checked, *hapax legomena* can be considered as a reliable indicator of productivity.

For the purpose of this study, I chose to employ the measure of "potential productivity" presented above, provided that its value is calculated at equal token numbers, and that the data are checked by manual inspection before measuring productivity.

3. The data

The empirical basis of this study is constituted by the RNC, a corpus of Russian created in 2003 by the Institute of Russian Language (Russian Academy of Sciences), which at the present moment contains over 600 million words. The RNC is one of the most important resources worldwide for the study of Russian and it is not only used by linguists, but also by teachers, students, writers, journalists, and in general by anyone interested in the Russian language. The RNC includes various subcorpora, the largest of which is the main corpus, a collection of texts representing standard Russian and containing both written texts (from the 18th century to the present day) and real-life Russian speech. Other subcorpora of the RNC are the parallel corpus, the poetry corpus, the dialectal corpus, the corpus of spoken Russian, the educational corpus, and others.[2] The analysis presented in this study is based on the main corpus, which contains texts belonging to different genres (fiction, drama, memoirs and biographies, journalism and literary criticism, scientific and popular scientific texts, instructional texts, religious and philosophical texts, technical texts, business and jurisprudence texts, and finally, letters and diaries), for a total of over 200 million words.

3.1. *Samo*-compounds in Russian

The prefixoid *samo*- ('self') in Russian can be attached to nouns and adjectives (mostly deverbal), and occasionally to verbs, creating compounds in which its meaning can vary. According to the Ožegov online dictionary (ozhegov-online.ru), the constituent *samo*- (which is an allomorph of the reflexive pronoun *sam* 'self', followed by the linking vowel -*o*-) has to be understood as the first element of compound words which may bear different meanings:

(1) orientation of something towards oneself, as in *samozaščita* ('self-defense')
(2) addressing oneself, e.g. *samouverennyj* ('self-confident')
(3) accomplishment of something without any external help, such as *samolečenie* ('self-medication')
(4) realizing something automatically, e.g. *samoventilacija* ('self-ventilation')
(5) autocracy, as in *samovlastie* ('autocracy')
(6) superlative, such as *samovažnejšij* ('the most important')

Of all these meanings, (6) must be left aside, because in this case *samo*- is the first constituent of compound adjectives in the superlative form and stands for *samyj* ('the most'). The meaning mentioned under (5) is quite

[2] See http://ruscorpora.ru/en/corpora-stat.html.

complicated too, because it is very often the case that these compounds are the result of a calque, usually from Ancient Greek (through Old Church Slavonic), and so it is not always easy to unpack the compound meaning, or at least we need to understand it by looking at the structure of the calqued word. As for the other cases, we could reduce the four meanings left to two main interpretations of *samo-*: one in which this element has a pronominal function with the meaning of 'self', which I represent as [samo]$_{P}$ (1 and 2 above), and one in which it has an adverbial function with the meaning of 'autonomous/autonomously, by oneself', represented as [samo]$_{A}$ (3 and 4 above):

a) [[*samo*]$_{P}$ X] 'self X' → *samozaščita* 'self-defense', *samouverennyj* 'self-confident'
b) [[*samo*]$_{A}$ X] 'autonomous/automatic X'[3] → *samolečenie* 'autonomous treatment', *samoventilacija* 'automatic ventilation'

So, in the case of [*samo*]$_{P}$, the action described by the underlying verb is addressed or oriented towards the speaker, whereas in the case of [*samo*]$_{A}$, the action is performed autonomously, automatically, by oneself. This distinction is consistent with earlier studies on reflexive compounds in Russian (Kibrik 2003) and in other languages (König 2011). The labels [*samo*]$_{P}$ and [*samo*]$_{A}$ that I use roughly correspond respectively to *REFL-derivacija* and *SAM-derivacija* in Kibrik's terminology (Kibrik 2003: ch. 15), and to "adnominal" (in construction with a reflexive marker) and "adverbial exclusive" in König's words (König 2011). Both scholars emphasize the importance of the focus in these compounds. König suggests that in compounds of the adnominal type, such as "self-irony", the focus is on the object, so the target is remarkable (we are usually ironic about other people and not about ourselves), while in compounds of the adverbial exclusive type, such as "self-loading", the focus is on the subject, so the agent is remarkable (it is remarkable that the subject

[3] Although most of the times in English the translation for such compounds would still be 'self X', I use instead the term "autonomous" to highlight the difference in meaning between the two patterns.

referent is both the source and the target of the action described by the underlying verb). In (1) and (2) one example[4] for each type is displayed.

(1) *I* ***samoironija*** *u Anjuty est', no èto* ***samoironija*** *osobaja: bez nenavisti k sebe – naoborot, s polnym prinjatiem sebja.* 'And Anjuta has **self-irony**, but it is a particular **self-irony**: without hate towards herself – on the contrary, with a full acceptance of herself.' [Natal'ja Zajceva. Anjuta Dlinnyjčulok // «Russkij Reporter», 2014]

(2) *No v gospitale pered otpravkoj, po-vidimomu, ne našlos' pistoleta, i on vynužden byl vooružit'sja* ***samozarjadnoj*** *vintovkoj.* 'But apparently before leaving the hospital he did not find the gun and he was obliged to arm himself with a **self-loading** rifle.' [Vasil' Bykov. Boloto (2001)]

Similarly to reflexive compounds in other languages, *samo*-compounds in Russian are usually characterized by componential and transparent semantics. Nonetheless, these compounds seem to exhibit various degrees of lexicalization, some of them appearing to be more lexicalized than others. Like productivity, lexicalization is another important notion for word-formation, and it has to be intended as "the whole process whereby an established word comes to diverge from the synchronically productive methods of word-formation" (Bauer 2001: 45). As already mentioned in previous works (see Fernández-Domínguez 2010), productivity and lexicalization seem to be (at least indirectly) interlinked and seem to be related to word frequency in different ways: lexicalized items show high frequencies, while items created through productive patterns are usually characterized by low frequencies. Although some of the compounds analyzed seem to show a high degree of lexicalization (see Section 5 for detail), intuitively the prefixoid *samo*- seems to be quite productive, since it is currently used to create neologisms that are not attested in dictionaries yet, such as *samoinkassacija* ('self-cash-in'), *samofil'tracija* ('self-filtration'), *samoapgrejd* ('self-upgrade'), and many

[4] All the examples are extracted from the RNC, http://ruscorpora.ru/.

more. However, in order to provide empirical evidence of the increasing profitability of the prefixoid *samo-* in the formation of compounds in Russian, it is necessary to perform a quantitative analysis based on a large number of data. But before moving to the analysis proper, we shall first focus on the corpus creation and pre-processing.

3.2. Corpus pre-processing

In order to observe the (presumably) increasing profitability of *samo-*compounds in Russian over time, I first created four subcorpora from the main corpus, each one representing a time period: the 18th century, the 19th century, the 20th century and the period that lasts from the beginning of the 21st century to the present day, as shown in Table 1. As can be observed, the four subcorpora differ in size, which is problematic if we want to apply the measure of potential productivity discussed in Section 2. However, we shall leave this problem aside for the moment and focus on the selection of *samo*-compounds from the subcorpora. If we simply search for *samo** in each subcorpus, we get the figures presented in the third column of Table 1 ("Before").

Table 1. Size of subcorpora (F) and tokens of *samo** before and after (N) pre-processing

Subcorpus	Size (F)	Tokens		
		Before pre-processing	After pre-processing	
(1) 1700–1799	3,715,941	(1,532)	789	(52%)
(2) 1800–1899	48,985,844	(27,948)	16,734	(60%)
(3) 1900–1999	120,166,543	(88,751)	59,922	(68%)
(4) 2000–2015	61,275,335	(46,979)	34,294	(73%)

However, the input data need to be corrected before applying the measure to calculate productivity, as pointed out in earlier studies on morphological productivity (Lüdeling et al. 2000; Evert & Lüdeling 2001;

Gaeta & Ricca 2006; Štichauer 2009). Indeed, the measure of potential productivity "does not yield the expected results unless the data is pre-processed according to a very good understanding of the morphological process in question" (Lüdeling et al. 2000: 57). What follows is a list of problems that have been addressed before carrying out the analysis:

a) all proper names (e.g. *Samodurov*, *Samochvalov*, *Samojlov*, etc.) were excluded from the analysis;[5]
b) adjectives in which *samo-* has a superlative function (e.g. *samonovejšij* 'the newest', *samoglavnejšij* 'the most important', *samoumnejšij* 'the most intelligent', and so on) were eliminated, because the morphological process at play is different from the one under investigation, as discussed in Section 3.1;
c) non-compounds (e.g. *samost'* 'the self') were eliminated. This is the case of baseless derivatives, as Gaeta & Ricca (2006: 74) define them;
d) orthographic variants (e.g. *samozabvenie/samozabven'e* 'self-oblivion', *samoljubie/samoljub'e* 'self-love', and so on) were fused, since they are variants of the same lexeme, the only difference being their use in different time periods;
e) for every derivational cluster, only the base word was kept, since the derived words are not representative of the productivity of *samo-*, but of the productivity of the whole compound. For example, the compound *samovar* ('samovar') is the source for other words, such as *samovarnyj* ('relative to samovar'), *samovarničat'* ('to drink tea sitting close to the samovar'), *samovarnik* ('merchant'), etc. If we considered all the words derived from *samovar*, we would not be investigating the productivity of *samo-*, but that of the whole compound *samovar*. This problem can be related to what Gaeta & Ricca (2006: 79) refer to as "inner derivational cycles" regarding suffixation in Italian. The authors opt for including in their analysis only the "outmost derivational cycle", that is to say, only those lexemes in

[5] All the pre-processing operations were carried out manually.

> which the suffix is attached last (for example, if we wanted to investigate the productivity of the suffix *-ment* in English, we would include the lexeme 'development', but not 'underdevelopment', which would count instead as a token of the prefix *under*-). Our case is somewhat different because we are dealing with a prefixoid, but in a similar way we could say that we consider only those compounds in which this element is attached last, or, at most, simultaneously with a suffix. So, for example, we take the adjective *samostojatel'nyj* ('independent') because there does not exist a verb such as **samostojat'*, nor a deverbal agent noun such as **samostojatel'*, but we do not take the adjective *samoletnyj* ('relative to aircraft') because we have the noun *samolet* ('aircraft'), from which the adjective is derived.

Once the pre-processing operations are carried out, the four subcorpora are reduced to the figures listed in the column "After" in Table 1.

For each subcorpus, every type and its corresponding token frequency was retrieved. The operation was carried out manually, because the corpus does not allow automatic extraction of all the types with *samo*-. This basically means that we need to scroll all the results of the search '*samo**' for each subcorpus in the RNC interface and transcribe in an spreadsheet file every new type we encounter. For each type, then, we look at the corresponding number of tokens in the subcorpus. The result of this operation is a very simple two-column spreadsheet file, in which the first column is filled with every new type encountered (e.g. *samovažnost'* 'self-importance', *samovar* 'samovar', *samovidec* 'eye-witness', and so on) and the second column with the number of tokens retrieved for each type in the subcorpus. The operation is repeated for each subcorpus. The figures obtained are summarized in Table 2, where N is the number of tokens, V the number of types, and V1 the number of *hapax legomena* with *samo*- for each subcorpus.

I now apply the measure of potential productivity presented in Section 2 (P = V1/N) to the data as they appear after the pre-processing. The results can be observed in the rightmost column of Table 2, where P stands for "potential productivity".

Table 2. Tokens (N), types (V), *hapax legomena* (V1) with *samo-* and potential productivity rate (P) for each subcorpus

Subcorpus	Tokens (N)	Types (V)	Hapax legomena (V1)	P
(1) 1700–1799	789	51	15	0.019
(2) 1800–1899	16,734	209	68	0.004
(3) 1900–1999	59,922	563	190	0.003
(4) 2000–2015	34,294	551	223	0.006

Contrary to our expectations, the results seem to show that the productivity of *samo*-compounds in Russian decreases over time (with just a slight recovery in the 21st century). However, these results are counterintuitive, and most likely an artifact of the productivity measure employed. Indeed, as mentioned above, Baayen’s measure of potential productivity is problematic when it is applied to differently-sized corpora (Lüdeling et al. 2000; Evert & Lüdeling 2001; Gaeta & Ricca 2006; Štichauer 2009; Efthymiou et al. 2012). This is due to the fact that the value of P is highly dependent on the sample size, because P is a function of N. In order to understand whether productivity increases over time or not, we need to calculate its value at equal token numbers, as shown in Section 4.

4. Applying LNRE models to calculate productivity over differently-sized corpora

The problem of comparing differently-sized corpora for measuring the productivity of a certain morphological process can be overcome in different ways: (a) by dividing the larger subcorpora into smaller pieces comparable to the shortest one; (b) by resorting to parametric statistical models of frequency distribution known as LNRE (Large Number of Rare Events) models (Baayen 2001). The first option proved to be unfeasible because the RNC interface does not allow specifying the size when creating subcorpora and the corpus cannot be entirely downloaded. For this reason, I decided to opt for the latter solution, also following other scholars’ proposals (Štichauer 2009; Efthymiou et al. 2012).

LNRE models, which allow extrapolating the expected values of types (V) and *hapax legomena* (V1) with a given affix for arbitrary values of tokens, are implemented in the package *zipfR* (Baroni & Evert 2014), a tool for lexical statistics in *R*, which is used for this study. The tool supports three classes of LNRE models: the Generalized Inverse Gauss Poisson model, the Zipf-Mandelbrot model and the finite Zipf-Mandelbrot model. The model chosen to carry out the present analysis is the Zipf-Mandelbrot model, which, as pointed out in previous works on morphological productivity (Štichauer 2009: 143), seems to perform best also with smaller samples. As such, the Zipf-Mandelbrot model can be used to estimate the expected values of types (EV) and *hapax legomena* (EV1) at equal token numbers (N) for each subcorpus. The token value is arbitrarily set at N = 25,000. The results of this estimate are summarized in Table 3. Once potential productivity is calculated at equal token values, the results meet the expectation that the productivity of the prefixoid *samo-* in Russian increased over time.

Table 3. Expected types (EV) and *hapax legomena* (EV1) (rounded to integers) and productivity rates for the four subcorpora calculated based on the Zipf-Mandelbrot model at the equal token values N = 25,000

	Expected		**Potential**
Subcorpus	**Types (EV)**	**Hapax legomena (EV1)**	**Productivity Rate (P)**
(1) 1700–1799	147	42	0.0017
(2) 1800–1899	234	66	0.0026
(3) 1900–1999	416	145	0.0058
(4) 2000–2015	487	190	0.0076

Figure 1 shows how the vocabulary size, that is, the number of types (plotted on the y-axis), increases over time (plotted on the x-axis). The expected number of types, estimated at the equal token value N = 25,000 (with 95% confidence intervals), are plotted at the middle of the time period (1750, 1850, 1950, 2007). As the number of types increases over time, the number of *hapax legomena* also increases (as shown in Table 3), and the productivity rates become higher.

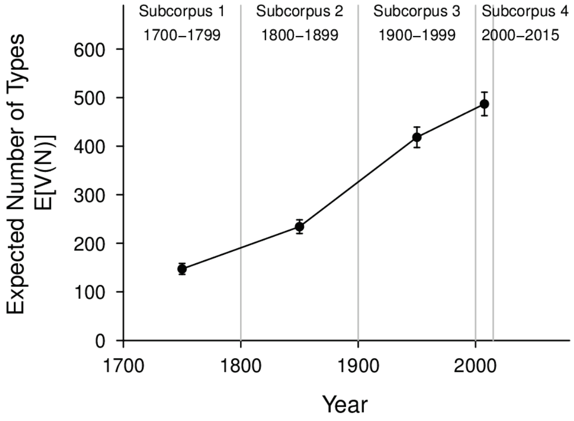


Figure 1. Expected types E[V(N)] with *samo-* in each subcorpus at the equal token value N = 25,000 (with 95% confidence intervals)[6]

5. Productivity vs. lexicalization

As mentioned in Section 3.1, productivity and lexicalization are two fundamental notions in word-formation, and they are somewhat interrelated. In particular, we saw that they behave in different ways with respect to word frequency: lexicalized items show high frequencies, while items created through productive patterns are usually characterized by low frequencies (cf. Fernández-Domínguez 2010). Again, as Aronoff & Anshen (1998: 245) put it, "the less productive a morphological pattern is, the more frequent on average its members will be". In fact, affix productivity is often associated with low-frequency words, while lexicalization is often believed to be caused by frequency of usage: a word which is frequently used is more likely to become lexicalized (Lipka 2002: 111; Bakken 2006: 107). In other words, if we assume that the formation of *samo*-compounds is a productive process in Russian, we should not expect a high number of lexicalized words. Indeed, if we take a closer look at the data we notice that very few words have a high

[6] I would like to thank Lukas Sönning for helping me with the chart design.

frequency in the corpus, while most words show low frequencies. The ten most frequent words for each subcorpus are shown in Table 4.

By looking at the most frequent items reported in Table 4, some interesting observations emerge. The largest part of these lexemes is not formed by adding the prefixoid *samo*- to a noun or an adjective, but the second member of the compound is rather a pure verbal root, as in (3), or a verbal root followed by a suffix, as in (4).

(3) *samovar* ('samovar') → *samo* + *var* (from *varit'* 'to boil')
samolet ('aircraft') → *samo* + *let* (from *letat'* 'to fly')
(4) *samoljubie* ('self-esteem') → *samo* + *ljub* (from *ljubit'* 'to love') + *i(j)-e*
samodel'nyj ('self-made') → *samo* + *del'* (from *delat'* 'to make') + *n-yj*

These word-formation patterns are indeed less productive in the creation of *samo*-compounds in Russian, while the most productive patterns are the following: [*samo* + N], exemplified by *samocenzura* ('self-censure'), [*samo* + A], exemplified by *samozarjadnyj* ('self-loading') and [*samo* + V], exemplified by *samoustanavlivat'sja* ('to self-center'), which also give rise to low-frequency words. What we can say, by looking at the data, is that there is a large number of low-frequency words formed through the productive patterns [*samo* + N], [*samo* + A], and [*samo* + V], while there is a small number of high-frequency words formed (not only, but also) through less productive morphological patterns. The high degree of lexicalization of these few high-frequency words is also demonstrated by the fact that they often give rise to word-formation series that also include compounds, which leads us to suppose that these lexemes are somewhat perceived as monomorphemic units, as exemplified in (5) and (6).

(5) *samolet* ('aircraft') → *samoletovoždenie* ('air navigation'), *samoletostroenie* 'aircraft construction'
(6) *samogon* ('moonshine') → *samogonovarenie/samogonokurenie* ('preparation of the moonshine')

Table 4. The ten most frequent items in each subcorpus and corresponding absolute frequency (N) and relative frequency per thousand words (ptw)

	Subcorpus 1			**Subcorpus 2**			**Subcorpus 3**			**Subcorpus 4**		
Rank	**Item**	**N**	**Ptw**	**Item**	**N**	**ptw**	**Item**	**N**	**ptw**	**Item**	**N**	**ptw**
1	*~ljubie*	177	224	*~ljubie*	2,356	141	*~let*	15,001	250	*~let*	10,106	295
2	*~deržec*	126	160	*~var*	2,212	132	*~stojatel'nyj*	6,87	115	*~stojatel'nyj*	5,818	170
3	*~vlastie*	58	74	*~stojatel'nyj*	1,693	101	*~var*	3,473	58	*~upravlenie*	2,461	72
4	*~vol'nyj*	57	72	*~zvanec*	984	59	*~ubijstvo*	2,963	49	*~ubijstvo*	993	29
5	*~zvanec*	53	67	*~dovol'nyj*	836	50	*~ljubie*	1,722	29	*~soznanie*	612	18
6	*~proizvol'nyj*	45	57	*~ubijstvo*	753	45	*~deržavie*	1,575	26	*~čuvstvie*	530	15
7	*~glasnyj*	44	56	*~otverženie*	580	35	*~upravlenie*	1,331	22	*~opredelenie*	449	13
8	*~praktičeskij*	35	44	*~bytnyj*	561	34	*~soznanie*	1,233	21	*~del'nyj*	434	13
9	*~deržavie*	21	27	*~deržavie*	437	26	*~del'nyj*	1,212	20	*~ljubie*	431	13
10	*~ubijstvo*	20	25	*~vol'nyj*	427	26	*~čuvstvie*	998	17	*~organizacija*	412	12

Therefore, within the same word-formation pattern (that of *samo*-compounds), we actually find some subpatterns, which exhibit vary- ing degrees of productivity. To provide empirical evidence for this assumption, I test the potential productivity measure on these different subpatterns using Subcorpus 4.[7] I created two subcorpora, one containing all the items that show the pattern [*samo* + X], that is to say [*samo* + N], [*samo* + A], and [*samo* + V], and the other one containing all the items in which *samo*- is attached to a verbal root (R), followed or not by a suffix (S), labelled [*samo* + XR (+S)]. Accordingly, the first group includes items such as *samocenzura* ('self-censure') or *samozarjadnyj* ('self-loading'), while the second group includes items such as *samolet* ('aircraft') or *samoljubie* ('self-esteem'). Quantitative results are summarized in Table 5.

Despite the fact that the two groups have similar sizes, I opted for calculating potential productivity at the equal token number of N = 25,000, estimating the number of types (EV) and *hapax legomena* (EV1) with the Zipf-Mandelbrot model, exactly as in Section 4. The results are reported in Table 5.

Table 5. Results for [*samo* + X] and [*samo* + XR (+S)]: Tokens (N), types (V), *hapax legomena* (V1), expected types (EV), expected *hapax legomena* (EV1) and potential productivity rate (P) at $N = 25{,}000$

				Expected		**Potential**
Construction	**Tokens (N)**	**Types (V)**	**Hapaxes (V1)**	**Types (EV)**	**Hapaxes (EV1)**	**productivity rate (P)**
[*samo* + X]	13,660	505	219	652	275	0.0110
[*samo* + XR (+S)]	20,634	46	4	46	5	0.0002

The results show that, as expected, the pattern [*samo* + X] is far more productive than the pattern [*samo* + XR (+S)]. It is in this second, less productive group that most of the lexicalized items of our sample occur,

[7] Subcorpus 4 is chosen for two main reasons: its size is in-between Subcorpus 2 and Subcorpus 3 (leaving aside Subcorpus 1, which is far smaller than the others), and it represents the language of the 21st century, so it might give us a better representation of the current trends in contemporary Russian.

which is in agreement with the theoretical assumption discussed above, that is, that productivity and lexicalization go into different directions: productive patterns give rise to a large number of low-frequency words, whereas lexicalized items are high-frequency lexemes mainly formed through less productive patterns.

If we now compare Figure 1 (showing the expected increase in the number of types with *samo*- over time) with Figure 2 (showing the expected increase in the number of types for the two patterns just discussed), we can observe that not only is there a difference in the productivity of the prefixoid *samo*- over time, but also that there is an even bigger difference in the productivity of *samo*- when the prefixoid is embedded in different constructions: as just discussed, the construction [*samo* + X] is far more productive than the construction [*samo* + XR (+S)].

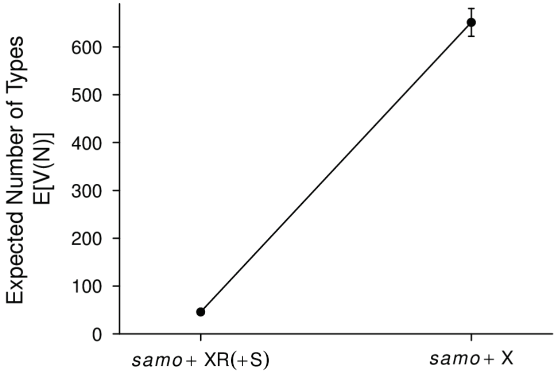


Figure 2. Expected types E[V(N)] with [*samo* + X] and [*samo* + XR (+S)] at the equal token value N = 25,000 (with 95% confidence intervals)

6. Discussion and conclusions

The analysis carried out throughout this paper provided empirical evidence for the increasing productivity of the prefixoid *samo*- in Russian over time. The research showed how quantitative methods can prove essential to exactly determine certain linguistic notions, such as that of morphological productivity. The initial hypothesis that *samo*-compounds in Russian have become more and more productive over time could thus be verified.

The impossibility of applying the measure of potential productivity to differently-sized corpora led us to resort to statistical methods of frequency distribution (LNRE models), which, on the basis of empirical data, allow estimation of the number of types and *hapax legomena* with a certain affix at arbitrary token values.

The results achieved showed that the morphological pattern under investigation is subject to a sharp rise in productivity, especially starting from the beginning of the 20th century. This might be (at least partially) due to the new terminology introduced following the developments in the fields of engineering and technology that took place during the 20th century. Examples (7) and (8) show how *samo*-compounds can be used to describe the functionalities of some newly created (or not yet created!) machines that perform particular actions in an automatic way.

(7) *V japonskom gorode Nagoja prošel Vsemirnyj čempionat po futbolu sredi* ***samochodnych, samobegajuščich*** *i* ***samoprygajuščich*** *avtomatov vsech razmerov i tipov.* 'In the Japanese city of Nagoya an international football championship was held between **self-walking**, **self-running** and **self-jumping** machines of all types and measures.' [Vo vsem mire // «Znanie - sila», 1998]

(8) *Skoree vsego imenno na ich osnove budut sozdavat'sja komp'jutery zavtrašnego dnja* ***samonastraivajuščiesja, samoremontirujuščiesja*** *i* ***samoobučajuščiesja*** *sistemy, struktura kotorych menjaetsja v zavisimosti ot rešaemoj zadači.* 'Most likely future computers will be built on their basis **self-adjusting**, **self-repairing** and **self-learning** systems whose function changes according to the problem being solved.' [Grigorij L'vov. Komp'juter strojat biologi // «Technika - molodeži», 1989]

Despite the conclusion that *samo*-compounds in Russian show an increasing productivity, it emerged that there is a small number of high-frequency lexicalized items that are usually formed through less productive patterns. Therefore, I shall conclude that the formation of compounds with the prefixoid *samo-* is a productive morphological process, but its productivity varies according to the different possible patterns. The less productive patterns are associated with a small number of lexicalized words, which exhibit a high token frequency in our sample.

One question not addressed in the present research, and which might be investigated in future research, concerns text types and registers, and their possible relation to productivity. As demonstrated in earlier studies (Efthymiou at el. 2012), the productivity of a certain morphological process may indeed vary across different text types and registers, so it might be interesting to enquire whether this concerns *samo*-compounds as well.

References

Aronoff, Mark & Frank Anshen. 1998. Morphology and the lexicon: Lexicalization and productivity. In: Andrew Spencer & Arnold Zwicky (eds.), *The handbook of morphology,* 237–247. Oxford: Blackwell.

Baayen, Harald R. 1992. Quantitative aspects of morphological productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1991,* 109–149. Amsterdam: Springer.

Baayen, Harald R. 1993. On frequency, transparency and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of morphology 1992,* 181–208. Dordrecht: Kluwer.

Baayen, Harald R. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, Harald. R. & Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29(5). 801–843.

Bakken, Kristin. 2006. Lexicalization. In: Keith Brown (ed.), *Encyclopedia of language and linguistics,* 106–108. Oxford: Elsevier.

Baroni, Marco & Stefan Evert. 2014. The zipfR package for lexical statistics: A tutorial introduction, http://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf (20 March, 2016).

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Booij, Geert. 2010. *Construction morphology*. Oxford: Oxford University Press.

Corbin, Danielle. 1987. *Morphologie dérivationelle et structuration du lexique*. Tübingen: Niemeyer.

Efthymiou, Angeliki, Georgia Fragaki & Angelos Markos. 2012. Productivity of verb-forming suffixes in Modern Greek: A corpus-based study. *Morphology* 22. 515–543.

Evert, Stefan & Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In: Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie & Shereen Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, 167–175. Lancaster: UCREL.

Fernández-Domínguez, Jesús. 2010. Productivity vs. lexicalization: Frequency-based hypotheses on word-formation. *Poznań Studies in Contemporary Linguistics* 46(2). 193–219.

Gaeta, Livio & Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.

Gaeta, Livio & Davide Ricca. 2015. Productivity. In: Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation: An international handbook of the languages of Europe*, 842–858. Berlin: Mouton de Gruyter.

Kibrik, Aleksandr E. 2003. *Konstanty i peremennye jazyka*. Saint-Petersburg: Aletejja.

König, Ekkehard. 2011. Reflexive nominal compounds. *Studies in Language* 35(1). 112–127.

Lipka, Leonhard. 2002. *English Lexicology: Lexical structure, word semantics and word-formation*. Tübingen: Narr.

Lüdeling, Anke, Stefan Evert & Ulrich Heid. 2000. On measuring morphological productivity. In Werner Zühlke & Ernst Günter Schukat-Talamazzini (eds.), *KONVENS-2000 Sprachkommunikation*, 57–61. Berlin: VDE.

Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin: Mouton de Gruyter.

Plag, Ingo. 2006. Productivity. In Bas Aarts & April McMahon (eds.), *The handbook of English linguistics,* 537–556. Oxford: Blackwell.

Štichauer, Pavel. 2009. Morphological productivity in diachrony: The case of deverbal nouns in *-mento, -zione* and *-gione* in Old Italian from the 13th to the 16th century. In Fabio Montermini, Gilles Boyé & Jesse Tseng (eds.), *Selected proceedings of the 6th Décembrettes,* 138–147. Somerville: Cascadilla Proceedings Project.

No matter how hard we try: Still no default plural marker in nonce nouns in Modern High German

Eugen Zaretsky and Benjamin P. Lange
University Hospital Frankfurt/Main and University of Würzburg

In the current article, 24 nonce nouns with or without rhymes in the German language are analyzed with respect to the distribution of plural allomorphs in the pluralizations of native speakers. The influence of several intralinguistic variables on the choice of plural markers is assessed: grammatical gender, word-final phonemes, classification of nonce words as those having or not having rhymes in German, plural markers of the rhyming real words, unusual orthography, final-obstruent devoicing, and the possibility of umlauting. The study aims to replicate the results presented in Marcus et al. (1995) by means of using the same test items, but with a different study design and a new sample of adult German native speakers ($N = 585$). The article emphasizes the methodological aspects and demonstrates the variability of findings depending on the chosen calculation method for binary, categorical, and linear regressions. Although the spectrum of possible results is quite broad, the present study, contrary to Marcus et al. (1995), allows to draw the conclusion that single-route models can better account for the distribution of plural markers than dual-route models.

1. Introduction

In the morphology of German, a wide range of pluralization patterns is utilized: apart from the plural markers *-e* (with or without umlaut), *-(e)n* (without umlaut), *-s* (without umlaut), *-er* (with umlaut), and umlaut alone, German nouns can be pluralized by zero markers, several markers borrowed from other languages (e.g. *Numerus* 'number' → *Numeri*) as well as various irregular forms with modifications of the stem (e.g. *Stadion* 'stadium' → *Stadien*) (Mugdan 1977). Further, several ungrammatical plural allomorphs can be found in spontaneous reactions to nonce nouns (Zaretsky et al. 2013a). Pluralization patterns are governed by a complicated constellation of intra- and extralinguistic factors, such as frequency of the plural markers and age of (preschool, but not adult) test subjects (Zaretsky & Lange 2014).

Almost all previous studies on plural acquisition worked with elicitation tasks and/or analyses of spontaneous speech (Clahsen et al. 1992; Szagun 2001; Korecky-Kröll & Dressler 2009). Studies with plausibility scales (Marcus et al. 1995; Korecky-Kröll et al. 2012) are scarce. Different study designs might naturally result in somewhat different findings, such as a higher percentage of zero forms in elicitation tasks compared to spontaneous speech (Clahsen et al. 1992). Due to the nature of elicitation tasks, namely a forced choice of only one plural form, they might be less sensitive to the subtle differences in the pluralization patterns in comparison with plausibility scales. In the latter case, test subjects can estimate the appropriateness of every plural allomorph, which, again, might result in different findings compared to elicitation tasks and spontaneous speech.

In the present study, the influence of the study design on its results was examined on the basis of Marcus et al. (1995). By means of plausibility scales these authors demonstrated a statistically significant preference for *s*-forms in comparison with all other plural markers in a sample of German adults. By contrast, in most other studies (e.g. Mugdan 1977; Wegener 1994; Elsen 2001; Szagun 2001), three plural forms dominated, namely *-s*, *-(e)n*, and *-e* (all three without umlaut), both in the answers of children and adults. The present study used a study design comparable to Marcus et al. (1995) – the same test population, namely adults, and the same nonce words – but with several modifications which were assumed to result in quantitatively or even qualitatively different findings: (i) elicitation tasks instead of plausibility scales, (ii) regressions instead of analyses of variance (ANOVA), (iii) a different classification of plural markers, and (iv) a much larger sample size.

Due to the peculiarities of plausibility scales (described above), it was expected that in the present study, with its standard elicitation tasks, test subjects would actively use those three plural allomorphs which were identified as the most frequent ones in Zaretsky et al. (2013a) and in many other studies, namely *-s*, *-(e)n*, and *-e*, instead of the preference for *-s* only. The use of these three plural markers, governed by such characteristics of German nouns as word-final phonemes and grammatical gender, would rather evidence single-route models of plural acquisition as opposed to dual-route models.

Generally speaking, proponents of dual-route models (Marcus et al. 1995; Clahsen 1999; Niedeggen-Bartke 1999) divide plural allomorphs into two groups, default and irregular ones. *-s* is considered to be the only representative of the first group, default ones (sometimes *-(e)n* is also classified as default, but rather as a result of its misinterpretation by preschool children). All other plural markers belong to the second group, irregular ones. The default plural marker is believed to be added in emergency cases, that is, when nouns do not evoke associations with the acquired vocabulary and are thus treated as new material that demands a special marker.

Single-route models – such as Natural Morphology (Wurzel 1984; Dressler et al. 1987) or Cognitive Morphology (Bybee 1985; Köpcke 1993) – do not subdivide German plural markers into two groups. Instead, they focus on the characteristics of plural allomorphs and nouns trying to find regularities behind pluralization patterns not in the dichotomy of the default plural vs. irregular ones, but, rather, in the frequency, perceptibility, productivity, (poly)functionality, and other characteristics of the plural allomorphs as well as in some characteristics of the pluralized nouns such as word-final phonemes.

Marcus et al. (1995) delivered statistical evidence for the dual-route models. The plural allomorph *-s* received the highest plausibility values in comparison with other plural markers in the unusual language material: nonce nouns having no rhymes in Modern High German as well as nonce nouns presented as names and borrowings, in comparison with "normal" language material, that is, nonce nouns having rhymes in Modern High German and also nonce nouns presented as real German nouns ("roots").

However, several methodological issues may be raised with regard to Marcus et al. (1995). To the best of our knowledge, these have not been commented on to date. First, the quality of the data should meet several requirements in order to be examined in an analysis of variance (ANOVA). The authors did not mention whether the data were normally distributed and whether homogeneity of variance was checked (which, however, is commonly done in psycholinguistic studies). Both assumptions are often violated in linguistic data. In addition, plausibility scales cannot always be treated as metrical, and there is no indication in Marcus et al. (1995) that the data were z-transformed. Furthermore, the sample (N = 48) used in Marcus et al. (1995) was far too small for an ANOVA. In case of the study design utilized by these authors (two binary independent variables), such an ANOVA would have required a sample of at least N = 128 ($\alpha = .05$, $1-\beta = .80$, medium effect size .25) according to the software *G*Power* 3.1 (Faul et al. 2009). A sample size of N = 48 must have resulted in a power of only .40 for a medium-sized effect. In several cases, a by-item study design with N = 24 (items) was used for ANOVA, which might have further underpowered the calculations.

Underpowered statistical tests are exposed to the danger of not only missing statistical significance, but also of overestimating the influence of some factors due to the flawed probability values that became statistically significant only by chance due to a small sample size. This has been shown for total scores of German language tests (speech comprehension, vocabulary, grammar, articulation) by Zaretsky & Lange (2015) in a series of retrospective analyses of data collected in several studies on the language test development (cf. Schüller 2015: 267).

High instability of results based on small samples can be exemplified by the following simulation. Twelve non-overlapping subsamples of equal size (N = 48, sample size used in Marcus et al.) were extracted from the sample used in the current article. For non-rhymes, the percentages of occurrences of respective plural markers varied considerably depending on the subsample, for instance 0–5% (out of all plural markers) for *-er* without umlaut, 11–20% for *-s*, 18–33% for *-(e)n*, and 36–52% for *-e* without umlaut.

None of the issues mentioned above can be considered absolutely critical, but their combination surely did not contribute to the reliability of the ANOVA results of Marcus et al. (1995).

The current article aims to demonstrate that a broad spectrum of statistically significant findings is possible even within the same methodological framework, depending on the calculation method. This phenomenon might help explain why numerous studies on the plural acquisition in German arrive at different results despite very similar study designs. For this purpose, this study contrasts several calculation methods for regressions.

Because other statistical methods are used in the current paper than in Marcus et al. (1995), the present study is not a mere replication of their results, but rather an analysis of the same test items and of a comparable test sample with a somewhat different study design. We believe that this design is more appropriate to answer the question of the original study, namely whether the distribution of the German plural markers can be better explained by single-route or dual-route models. We consider such a re-analysis necessary because even after decades of fierce debates and mutual criticism, proponents of both single-route and dual-route models have not arrived at any clear conclusion apart from declaring that the arguments of their respective counterparts must be wrong (for an overview, see Clahsen 1999; Hahn & Nakisa 2000).

We hypothesized that German adults would prefer plural markers *-s*, *-(e)n*, and *-e* (all three without umlaut), both with rhyming and non-rhyming nonce nouns; the first one because of its high compatibility with borrowings including many "non-rhyme" neologisms (i.e. neologisms without clear phonotactic analogies in German) and with word-final phonemes, the latter two because of their high frequency in Modern High German (Zaretsky et al. 2011). These three plural markers are characteristic of the answers of German preschoolers in nonce words tasks (Zaretsky et al. 2013a, 2013b). The assumption that adults would stick to the same pluralization patterns does not mean that they are not capable of analyzing unknown language material in more detail compared to children, but the test items chosen by Marcus et al. (1995) deliver very few cues on possible plural forms, which may force adults to use the simplest pluralization strategies. We also hypothesized that adults would choose the same three plural markers significantly more often for non-rhymes (test items without rhymes in Modern High German) than for rhymes (test items having such rhymes), because any associations with semantics and phonology are missing in case of non-rhymes. This would contradict the results of Marcus et al. (1995) and would rather support the single-route models.

After a brief outline of methodological issues (Section 2), we present several calculation methods (Section 3) to demonstrate that results (regarding associations between intra-/extralinguistic factors and the preference for certain plural markers) depend to a certain extent on such factors as the independent variables included in the generalized linear mixed models. In Section 4, results are discussed in terms of single-route and dual-route models and under consideration of methodological issues that may influence the outcome of calculations.

2. Methods

Test subjects were 585 adult German native speakers (age range 18–96 years, median 24); 207 males (35.4%), 369 females (63.1%), and nine participants with unknown sex (1.5%). They were recruited mostly at universities of the German state of Hesse during the years 2011–2013.

All 24 nonce words – twelve rhymes and twelve non-rhymes (see Appendix) – from Marcus et al. (1995) were presented to the participants in written form as real German nouns (cf. "roots" in the Marcus et al. study) and then compared with respect to the distribution of plural allomorphs. The rhymes were supposed to elicit clear associations with widely used real German nouns (e.g. nonce noun *Pind* → *Pinder* following the model of *Kind* → *Kinder* 'children'). Information on the most frequent types and, for comparison, tokens associated with the test items can also be found in the Appendix. The presentation of test items was not randomized and no filler items were used. The participants produced written responses.

Instead of a plausibility scale, participants were asked to actively produce plural forms because the chosen plural allomorph is obviously the most plausible one for the test subjects. We consider ordinal plausibility scales as not very appropriate for the analysis of internalized pluralization strategies because often one can use several plural markers of a nonce word depending on personal extra- and intralinguistic associations, priming, creativity, and motivation. In the present study, test subjects had to decide in favor of a single plural form in the production tasks. The analysis of only one plural form per item might make the results more reliable because it excludes numerous other forms which would never be actively produced by the test subject and which would be unequivocally rated acceptable even if they are only marginally acceptable.

Since the grammatical gender in Marcus et al. (1995) was dichotomized into feminine and non-feminine (because there are hardly any differences in the use of plural allomorphs between nouns of masculine and neuter gender), our test items were also presented as two gender groups. We expected that adults would use *-e* and probably also *-er* with non-feminine gender and *-(e)n* with feminine gender, because these associations are more or less clearly represented in Modern High German (Zaretsky et al. 2013b; cf. Wegener 1994). Following Marcus et al. (1995), the gender shift was applied, that is, nouns presented as feminine ones to one half of the test subjects were presented as non-feminine ones to the other half. The gender shift helps to control whether dichotomized gender influences the choice of plural allomorphs.

The following characteristics of the test items were chosen for the analysis:

- umlauting in the test items (that is, whether vowels can be subject to umlauting during pluralization),
- dichotomized grammatical gender (feminine vs. non-feminine nouns),
- word-final phonemes without any categorization,
- rhyme vs. non-rhyme,
- the most probable and second most probable associations with plural markers of real words for rhymes (types, according to the calculations by Ruoff 1981, cited in Marcus et al. 1995, and tokens; Institut für Deutsche Sprache 2009), e.g. *Pind* → *Pinder* following the model of *Kind* → *Kinder* ('children') (see Appendix),
- usual or unusual orthography (some of the items used by Marcus et al. 1995 contained rare grapheme combinations such as <hk> or <hf>),
- final-obstruent devoicing (a systematic devoiced pronunciation of voiced obstruents).

The relevance of most of the chosen characteristics for the distribution of plural markers has already been shown in previous studies (Mugdan 1977; Fakhry 2005). Among the characteristics of test subjects, only their age and sex were analyzed in the current study.

Several multivariate methods were applied to examine the influence of the characteristics of test items and test subjects (see above) on the distribution of plural allomorphs. For a dichotomized or categorical classification of plural allomorphs, an ANOVA – the method of choice in Marcus et al. (1995) – cannot be conducted. Instead, several calculation methods for regressions were chosen for comparison.

First, categorical regressions with classifications of plural allomorphs as dependent variables were calculated (Tutz 2011). This method can be considered comparatively simple because it does not differentiate between fixed and random factors. Categorical regression mirrors conventional multiple regression with the added property that this technique can also accommodate nominal variables (in this case classifications of German plural markers). It quantifies categorical data by assigning numerical values to the categories, resulting in an optimal linear regression equation for the transformed variables.

The regressions were calculated in three variants, namely with three different dependent variables for a subsequent comparison:

- ALL1: a detailed description of umlauting in the plural allomorphs: umlaut, *-er*, *-er* with umlaut, *-(e)n*, *-(e)n* with umlaut, *-s*, *-s* with umlaut, *-e*, *-e* with umlaut,
- ALL2: a less detailed description of umlauting (a classification used in Marcus et al. 1995): umlaut, *-er*, *-er* with umlaut, *-(e)n*, *-s*, *-e*, *-e* with umlaut,
- ALL3: plural allomorphs of Modern High German: umlaut, *-er*, *-(e)n*, *-s*, *-e*, *-e* with umlaut.

Plural allomorphs analyzed in ALL2 and Marcus et al. (1995) neither correspond completely to those applied in Modern High German (*-er* without umlaut is ungrammatical), nor do they represent in full detail possible combinations of plural markers with umlaut (ungrammatical combinations of *-s* and *-(e)n* with umlaut were not taken into account). The consideration of umlauting with only one plural marker, *-e* (*-e* with umlaut vs. *-e* without umlaut), in ALL3 is explained by the absence of separate rules for umlauting in case of other plural markers. Whereas *-er* always demands umlauting and both *-s* and *-(e)n* always forbid it, *-e*, as the only German plural allomorph, can occur either with or without umlauting, which motivates the subdivision into two different plural allomorphs (Mugdan 1977). In ALL1–ALL3, ungrammatical plural markers, such as *-s* with umlaut, were included.

For all characteristics of test items that yielded significant results in the categorical regressions, additional analyses of their associations with the plural markers were carried out in cross-tables to describe the results in terms of percentages.

Second, the influence of the same factors was analyzed for each plural allomorph separately in binary logistic regressions (e.g. *-s* vs. all other plural markers). This analysis can be considered comparatively sophisticated due to the differentiation between fixed and random factors (cf. Korecky-Kröll et al. 2012) within a generalized linear mixed model (GLMM). Dichotomized classifications of plural markers served as dependent variables, test subjects and test items as random factors (random intercepts), and their characteristics as fixed factors. A spectrum of possible results was demonstrated by varying the choice of factors. No calculations were feasible for umlaut due to its low frequency ($N = 14$).

Furthermore, binary logistic regressions calculated in the "long" data design (N = 14,040 answers of test subjects) were compared to the linear regressions in the "broad" by-item design (N = 24 items), both calculated within GLMM. Due to the lack of space, only significance values are reported, without additional information on coefficients, explained variance, etc.

According to the hypothesis, plural markers *-s*, *-(e)n*, and *-e* were preferred to other plural markers with rhymes and, even more so, with non-rhymes. This was additionally analyzed by Wilcoxon and *t*-tests for two-paired groups. Total scores of usage of respective plural markers with 12 rhymes and 12 non-rhymes (dependent variable) were paired for each test subject, which resulted in a total of six Wilcoxon tests and six *t*-tests. All statistical analyses were carried out in *IBM SPSS 21* (IBM 2012).

Table 1. Regressions with categorical and dichotomous classifications of plural allomorphs as dependent variables, characteristics of test items and subjects as independent variables.

	Categorical regressions			**GLMM (ALL3): Binary logistic regression with random factors for test items and subjects: Model 1**				
Factor	ALL1	ALL2	ALL3	*-s*	*-(e)n*	*-er*	*-e*	*-e* + umlaut
Rhyme vs. non-rhyme	*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Grammatical gender	**	***	***	*	***	*	***	*ns*
Gender shift	***	***	***	*	*ns*	*ns*	*ns*	*ns*
Word-final phoneme	***	***	***	*ns*	*	*ns*	*ns*	*ns*
Plural marker of high-frequency rhymes								
Type frequency: Rank 1	***	***	***	*ns*	*ns*	*ns*	*ns*	*ns*
Type frequency: Rank 2	***	***	***	*ns*	***	*ns*	*	**
Token frequency: Rank 1	***	***	***	—	—	—	—	—
Token frequency: Rank 2	***	***	***	—	—	—	—	—
Umlauting possible?	***	***	*	*ns*	*ns*	*	*ns*	***
Usual vs. unusual orthography	**	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*
Final obstruent devoicing	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Test items	***	***	***	*ns*	*ns*	*ns*	*ns*	*ns*
Test subjects	***	***	***	***	***	***	***	***
Age	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*
Sex	—	—	—	—	—	—	—	—

Table 1. (continued)

	GLMM (ALL3): Binary logistic regression with random factors for test items and subjects: Models 2–12																					
	M2		M3		M4		M5		M6		M7		M8		M9		M10		M11		M12	
Factor	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*	*-s*	*-(e)n*
Rhyme vs. non-rhyme	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Grammatical gender	*	***	*	***	*	***	*	***	*	***	*	***	*	***	*	***	*ns*	***	*	***	*ns*	***
Gender shift	*	*ns*	*	*ns*	*	*ns*	*	*ns*	*	*ns*	*	*ns*	*	*ns*	*	*ns*	*ns*	*ns*	*	*ns*	*	*ns*
Word-final phoneme	*ns*	*ns*	*ns*	***	*ns*	***	*ns*	***	*ns*	***	*ns*	*	*ns*	*	*ns*	**	*ns*	**	*ns*	*	*ns*	*ns*
High-frequency rhymes																						
Type freq. Rank 1	*ns*	*ns*	—	—	—$^{\mid}$	—$^{\mid}$	*ns*	*ns*	—	—	*ns*$^{\mid}$	*ns*$^{\mid}$	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*$^{\mid}$	*ns*$^{\mid}$	*ns*	*ns*
Type freq. Rank 2	*ns*	*ns*	—	—	—$^{\mid}$	—$^{\mid}$	—	—	—	—	*ns*$^{\mid}$	***$^{\mid}$	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*$^{\mid}$	***$^{\mid}$	*ns*	*ns*
Token freq. Rank 1	*ns*	*ns*	*ns*	*ns*	*ns*$^{\mid}$	*ns*$^{\mid}$	*ns*	***	*ns*	***	—$^{\mid}$	—$^{\mid}$	—	—	*ns*	*ns*	*ns*	*ns*	—$^{\mid}$	—$^{\mid}$	—	—
Token freq. Rank 2	*ns*	*ns*	*ns*	***	*ns*$^{\mid}$	***$^{\mid}$	—	—	*ns*	*ns*	—$^{\mid}$	—$^{\mid}$	—	—	*ns*	*ns*	*ns*	*ns*	—$^{\mid}$	—$^{\mid}$	—	—
Umlauting possible?	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
(Un)usual orthography	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*$^{\mid}$	*ns*$^{\mid}$	*ns*	*ns*
Final devoicing	*ns*	*ns*	*ns*	***	*ns*¦	***¦	*ns*	*ns*	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*¦	*ns*¦	*ns*	*ns*
Test items	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	***	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	***$'$	***$'$
Test subjects	***	***	***	***	***	***	***	*ns*	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Age	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Sex	—	—	—	—	—	—	—	—	—	—	—	—	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	—	—	—	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, *ns* = not significant

ALL1: umlaut, *-er*, *-er* with umlaut, *-(e)n*, *-(e)n* with umlaut, *-s*, *-s* with umlaut, *-e*, *-e* with umlaut

ALL2: umlaut, *-er*, *-er* with umlaut, *-(e)n*, *-s*, *-e*, *-e* with umlaut

ALL3: umlaut, *-er*, *-(e)n*, *-s*, *-e*, *-e* with umlaut; $^{\mid}$ = nested terms within variable "rhymes/non-rhymes“, ¦ = nested terms within variable "word-final phonemes", $'$ = variable "test subjects" as a fixed factor (other occurrences of this variable being a random factor).

3. Results

Table 1 gives an overview of categorical regressions (without random and fixed factors) and binary logistic regressions (with random and fixed factors). In categorical regressions, three classifications of plural allomorphs were compared with respect to the influence of intra- and extra-linguistic factors on the distribution of plural allomorphs. In binary logistic regressions, only ALL3 was utilized due to limitations of space.

Among binary logistic regressions, model 1 can be considered the "standard" or the "best" one in terms of information criteria and other factors (Baayen 2008) in comparison with all other regressions reported in Table 1. Therefore, in model 1, all plural allomorphs were examined as dependent variables. In models 2–12, to demonstrate the variability of results, constellations of independent variables were tried out only for plural allomorphs *-s* and *-(e)n*.

Apart from different constellations of test variables (e.g. inclusion or exclusion of sex of test subjects), variability of results in Table 1 can be traced back to the use of nested terms (models 4, 7, 11: associations with real words vary only within rhymes; models 4, 11: final-obstruent devoicing varies within word-final phonemes; model 11: unusual orthography varies within non-rhymes) and classification of test items either as a random factor (models 1–11) or a fixed factor (model 12). Although unusual by conventional standards, the classification of test items as a fixed factor might make sense because those were not freely chosen but taken over from Marcus et al. (1995). In a possible replication study, the same test items would have to be chosen again.

Multinomial regressions within GLMM, with all plural markers taken together as a dependent variable, were tried out as well but did not yield acceptable results in terms of model stability and information criterion quality.

According to categorical regressions (Table 1), all chosen variables apart from age and sex yielded statistically significant results at least once. Next, characteristics of the test items were analyzed by means of cross-tables in respect to the tendencies in the distribution of plural markers (ALL3).

First, no evidence was found that *-s* dominated with non-rhymes in comparison to other plural markers. In ALL3, among non-rhymes, percentages of the most actively used plural allomorphs were 16% (out of all plural allomorphs used with non-rhymes) for *-s*, 28% for *-(e)n*, and 47% for *-e* (cf. 7% for *-e* with umlaut, 3% for *-er*, 0% for umlaut). Among rhymes, the percentage of *-s* out of all plural allomorphs used was even smaller, namely 8% (cf. 20% for *-(e)n*, 45% for *-e*, 15% for *-e* with umlaut, 12% for *-er*, 0% for umlaut). The plural markers *-er*, umlaut, and *-e* with umlaut taken together were used more often with rhymes (75% out of all their uses) than with non-rhymes (25%), whereas the proportions of the markers *-s*, *-(e)n*, and *-e* taken together were more equally distributed in rhymes (45%) and non-rhymes (55%), being more frequent in the latter group.

Table 2 compares the frequencies obtained in this study with the plausibility values from Marcus et al. (1995) (ALL2 in contrast to ALL3 in the previous paragraph). Whereas according to the Marcus et al. (1995) data *-s* was the most plausible plural marker with non-rhymes, our data indicate that not *-s* but *-e* without umlaut dominated in these plural forms. As far as rhymes are concerned, data of both studies demonstrated the preponderance of *e*-forms (without umlaut), however, followed by *s*-forms in the Marcus et al. (1995) study and followed by *en*-forms in our study.

Table 2. Raw frequencies and percentages (out of all plural markers used with the respective category: rhymes vs. non-rhymes) of plural markers used with rhymes and non-rhymes in this study, compared to the plausibility scales in Marcus et al. (1995), with 5 meaning "perfectly natural" and 1 meaning "perfectly unnatural"

Plural markers (ALL2 classification)	**Rhymes**			**Non-rhymes**		
	This study		**Marcus et al. (1995)**	**This study**		**Marcus et al. (1995)**
	N	%		N	%	
umlaut	11	0	1.5	3	0	1.5
-er with umlaut	392	6	1.7	169	2	1.9
-er without umlaut	573	8	2.2	179	3	2.5
-(e)n with or without umlaut	1,375	20	2.6	1,882	27	2.4
-e without umlaut	3,086	44	3.8	3,221	46	3.4
-e with umlaut	1,006	14	2.8	440	6	2.7
-s with or without umlaut	519	8	3.5	1,074	15	3.8
Total	6,962	100		6,968	100	

Second, nouns of feminine and non-feminine gender demanded different plural allomorphs. With nouns of feminine gender, *-(e)n* was used more frequently (19% of all plural markers with non-feminine gender vs. 33% with feminine gender); with nouns of non-feminine gender, *-e* (49% vs. 41%) and *-er* (9% vs. 3%) were comparatively frequent. In case of *-s* (13% vs. 11%) and *-e* with umlaut (11% vs. 10%), the percentages were almost identical.

Third, gender shift resulted in different distributions of plural allomorphs, although *-e* dominated in both halves of the sample (46–47% of all plural allomorphs), followed by *-(e)n* with 22–26%.

Fourth, differences in the distribution of plural allomorphs also depended on word-final phonemes. For instance, *-e* occurred in 47% of cases after /f/ (that is, in the other 53%, other plural markers were chosen with this word-final phoneme), *-s* was most closely associated with the word-final phoneme /ŋ/ (18%), *-e* plus umlaut with /n/ (36%), umlaut with /r/ (1%), *-(e)n* with /r/ (49%), and *-er* with /x/ (26%).

Fifth, there was a certain association between nonce words and their existing rhymes (types) in respect to the choice of plural markers. However, the preference for the plural marker of the real words varied considerably, from only 9% in case of *-(e)n* (that is, out of all plural forms of the respective nonce word, 9% corresponded to the one in the rhyming real words, namely *-(e)n*) to 45% in case of *-e* with umlaut. The second most frequent associations also demonstrated varying percentages of the correspondences, from 9% in case of *-er* to 63% in case of *-e*. The same is valid for rhymes as tokens.

Sixth, for test items subject to umlauting (e.g. *Bnaupf*), plural allomorphs with umlaut were chosen more often than for other items (e.g. *Pind*). In fact, only one umlaut was registered in the items without vowels subject to umlauting, and this single case resulted from an item deformation (*Bneik* → *Bnäuke*).

Seventh, test items with unusual orthography, *Fnöhk*, *Bnöhk*, *Fnähf*, and *Pnähf*, had a higher percentage of the plural markers *-s* (14% of all plural markers with these four items vs. 11% with other items), *-e* (55% vs. 45%), and *-(e)n* (28% s. 23%). The plural markers with umlaut could not occur in the four items with unusual orthography at all because these already contain umlauting. The plural marker *-er* was hardly used with these four nouns (3% vs. 8%).

Lastly, final-obstruent devoicing was associated with a higher frequency of the plural markers *-er* (17% vs. 5% of all plural markers), *-e* with umlaut (15% vs. 10%), and a lower frequency of *-(e)n* (16% vs. 26%).

Apart from binary logistic regressions (Table 1), linear regressions are sometimes used for the same purpose. For this type of analysis, a "broad" data design is utilized (N = 24 test items) instead of a "long" design (N = 14,040 answers). In this case, characteristics of test subjects such as age and sex cannot be accounted for. In addition, since grammatical gender varied in the current study, it had to be excluded. Total scores of uses of the respective plural marker per item served as dependent variables. Results of linear regressions with the remaining independent variables, under consideration of fixed and random factors, are presented in Table 3 in comparison with binary logistic regressions calculated in the "long" data design with the same factors (ALL3).

Table 3. Regressions calculated within generalized linear mixed model, with characteristics of test items as independent variables (fixed factors): (1) linear regressions in the "broad" data design (N = 24 test items), with test items as random factors vs. (2) binary logistic regressions in the "long" data design (N = 14,040 answers), with test items and test subjects as random factors

	-s		*-(e)n*		*-e*		*-e* + umlaut		*-er*	
Factors	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Rhyme vs. non-rhyme	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Word-final phoneme	*	*ns*	***	*	*	*ns*	**	*ns*	*ns*	*ns*
High-frequency rhymes										
Type frequency: Rank 1	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Type frequency: Rank 2	*ns*	*ns*	***	***	**	*	**	**	*ns*	*ns*
Umlauting possible?	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	***	*	*
Usual/unusual orthography	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Final obstruent devoicing	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Test items	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*
Test subjects	—	***	—	***	—	***	—	***	—	***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$, *ns* = not significant

According to Wilcoxon tests for two paired groups, *-s*, *-(e)n*, and *-e* taken together occurred more frequently than other plural markers both in rhymes (Z = -19.62, p < .001, N = 585, mean/M = 8.51, standard deviation/SD = 2.64 vs. M = 2.06, SD = 2.14) and in non-rhymes (Z = -21.12, p < .001, N = 585, M = 10.56, SD = 1.82 vs. M = 1.08, SD = 1.44). Other results of the Wilcoxon tests are presented in Table 4. Exact *p*-values calculated with the Monte Carlo method are reported.

It is noteworthy that a two-tailed calculation would have resulted in a merely marginally significant *p*-value for the plural marker *-e* (p = .070) instead of p = .036 hidden behind one asterisk in Table 4. If one-tailed *t*-tests were used instead of Wilcoxon tests, the result of *e*-forms would remain marginally significant (p = .057), other differences being highly significant (all *p*s < .001). However, none of the metrical variables was normally distributed according to the Kolmogorov-Smirnov test (all *p*s < .001, all *N*s = 585), which makes the use of *t*-tests questionable.

Table 4. Wilcoxon tests (*Z*), mean values, and standard deviations (in brackets) of plural allomorphs used with rhymes and non-rhymes; *N*s = 585

	-s	*-(e)n*	*-e*	*-e* + umlaut	*-er*
Rhymes	0.89 (1.67)	2.35 (2.48)	5.28 (2.82)	1.72 (1.92)	1.36 (1.60)
Non-rhymes	1.84 (2.50)	3.22 (3.50)	5.51 (3.40)	0.75 (1.06)	0.32 (1.06)
Z	-9.97***	-6.31***	-1.80*	-11.24***	-13.38***

Note. * p < .05, ** p < .01, *** p < .001

4. Discussion

According to Marcus et al. (1995), a preference for *s*-forms on a plausibility scale with non-rhyming nonce nouns in comparison with rhyming ones suggests a default status of *-s* with unusual language material. In the present study, the same test items – 24 nonce nouns rhyming or not rhyming with real German nouns – were used in order to investigate the findings of Marcus et al. from a different perspective. Instead of a plausibility scale, German native speakers were asked to produce plural forms of nouns presented in the singular form. It was expected that instead of using *-s* as the default plural marker and also instead of a clear preference for *-s* in the plural forms, adults would utilize the same three plural markers typical of preschoolers, namely *-s* (as the most compatible one in phonotactic respects), *-(e)n*, and *-e* as the most frequent ones in Modern High German (ALL3; cf. Köpcke 1993; Wegener 1994).

Indeed, the most important finding of the current study was that German adults preferred the plural markers *-s*, *-(e)n*, and *-e* both with rhymes and, especially, with non-rhymes compared to *-er*, umlaut, and *-e* with umlaut. However, differentiation between rhymes and non-rhymes yielded statistically significant results only in one (categorical) regression and in the univariate tests. In more sophisticated analyses, regressions with fixed and random factors, the significance of this differentiation was not identifiable at all, which might indicate that *-s*, *-(e)n*, and *-e* dominated in plural forms irrespective of presence or absence of associations with existing German nouns (apart from those associations that were assessed in other independent variables).

According to Marcus et al. (1995), only the plausibility of the "regular" plural marker *-s* with nonce words was higher for non-rhymes than for rhymes (means 3.8 vs. 3.5, with 5 meaning "perfectly natural" and 1 meaning "perfectly unnatural"), whereas the plausibility of *-e* and *-(e)n* was lower (3.8 vs. 3.4 and 2.6 vs. 2.4, respectively). The plausibility of umlaut remained on the same level for both word groups. The values of *-er* were somewhat higher for rhymes than for non-rhymes both with and without umlauting (1.9 vs. 1.7 and 2.5 vs. 2.2, respectively). *-e* with umlaut was less plausible with non-rhymes, although the difference was minimal (2.8 vs. 2.7).

However, if one measures the plausibility not by means of a plausibility scale but by means of active pluralizations, as was done in the current study, other tendencies emerge in the same "roots" (i.e. nonce words presented as real German words). Not only the frequency of *-s*, but also the frequency of *-(e)n* and *-e* were higher for non-rhymes than for rhymes. The plural marker *-s* occurred less often than *-(e)n* and *-e* with non-rhymes as well as less often than *-(e)n*, *-e*, *-e* with umlaut, and *-er* with rhymes, which can hardly be expected of the plural marker considered as the only default one.

The plural allomorph *-e* without umlaut made up about a half of all pluralizations both with rhymes and non-rhymes but occurred only 2% less often with the former than with the latter (45% vs. 47% of all pluralizations; ALL3). At first sight, this difference might appear marginal and negligible in spite of its statistical significance. However, one can hardly call this result qualitatively different from those for *-s* (8% vs. 16% of all pluralizations) and *-(e)n* (20% vs. 28%), since each of them yielded only 8% more pluralizations with non-rhymes than with rhymes, that is, only 6% more than in case of *-e* without umlaut.

The choice of test items by Marcus et al. (1995) is of special interest in this respect. According to the type frequency list by Ruoff (1981), the reference used by Marcus et al. (see Appendix), rhyming real German nouns demanded *-e* in nine cases out of 24, *-e* with umlaut in eight cases, *-(e)n* in only one case, *-er* in four cases, *-s* in zero cases (and two further items had no second most frequent rhymes). It is highly probable that if plural markers in the rhyming nouns had been distributed equally, instead of demanding *-e* with or without umlaut in most cases, then the difference in *e*-uses between rhymes and non-rhymes would have been larger, and the difference in *s*- and *en*-uses smaller.

The three most frequently used plural markers in the current study, *-s*, *-(e)n*, and *-e*, have already been described as the most productive ones in nonce words pluralized by German children, with preference for *-s* in linguistically more advanced groups and preference for *-(e)n* in linguistically weaker groups (Mugdan 1977; Zaretsky et al. 2013c). Obviously, German adults make use of the pluralization strategies comparable to those of German children.

Although the dependence of the choice of plural markers on the inter-individual characteristics of the study participants was not the subject of this study, the fact that such differences actually existed was shown by a statistically significant *p*-value of the corresponding random factor in almost all regressions. The age of the test subjects yielded a statistically significant result as well, but only in one case, *-e* with umlaut (fewer such plural markers in the answers of older participants), and with a coefficient confidence interval reaching the level of -0.002, that is, a level extremely close to zero. Sex was tried out only marginally and did not yield any significant results.

The significant result of word-final phonemes demonstrates that German adults retrieve information on the possible plural marker on the basis of some kind of frequency analysis of compatibility of word-final consonants and plural allomorphs. It should be taken into account that word-final phonemes are, on their part, more or less closely linked to other factors such as grammatical gender (Zaretsky et al. 2013b).

We cannot rule out the possibility that the presentation of the test items in a written form may have influenced the choice of plural markers due to the misinterpretation of the word-final phonemes in such items as *Pund* (/...t/) where so-called final obstruent devoicing occurs, that is, voiced obstruents become voiceless in word-final position. Indeed, an influence of this factor was found in some regressions. A further influence of orthography could have been expected in the test items *Fnöhk, Bnöhk, Fnähf,* and *Pnähf*: Combinations of graphemes <hk> and <hf> occur rarely and might have disoriented some test subjects in the choice of the plural allomorphs. Again, the influence of this variable on the choice of plural markers turned out to be statistically significant in some regressions. However, both factors were not significant in the "best" regression (model 1 in Table 1), a relatively conservative binary logistic regression calculated within GLMM, so it is up to the researcher in such cases which result should be reported.

Although Modern High German differentiates between nouns of masculine and neuter gender, with somewhat different pluralization patterns, the current study followed the design of Marcus et al. (1995) with respect to non-differentiation between these two genders. It was predicted that *-(e)n* would occur more often with nouns of feminine gender, whereas *-e* and probably also *-er* would be used with nouns of non-feminine gender because such tendencies are typical of Modern High German (Mugdan 1977; Zaretsky et al. 2011; Zaretsky & Lange 2014). These tendencies were indeed found in cross-tables and confirmed by very stable significance values in the regressions (Table 1). No other factor delivered as statistically reliable results as those for grammatical gender. Significant results of the gender shift in some regressions obviously also show that test subjects sometimes tended to apply different pluralization strategies depending on the gender of the same test items.

Different categorizations of plural markers can result in somewhat different findings. In Table 1, three different categorizations of plural allomorphs (ALL1–ALL3) were compared in the categorical regressions with the same independent variables. The most detailed classification of plural allomorphs (ALL1) delivered more significant results than the least detailed one (ALL3). But the results were neither contradicting nor qualitatively different, except that the final obstruent devoicing yielded a highly significant result in ALL2, but no significant results in ALL1 and ALL3. Hence, the classification of *-s* and *-(e)n* as containing or not containing umlaut turned out to be more or less superfluous, probably due to a low number of ungrammatical forms in the answers of adults.

A non-differentiation between *-e* with and without umlaut that is also sometimes found in studies on pluralization in German (e.g. Spreng 2004) would have had more far-reaching consequences. Because the plural marker *-e* (without umlaut) was used more often with non-rhymes compared to rhymes, a classification of *-e* with and without umlaut as one plural marker would have resulted in a significantly higher proportion of *e*-pluralizations in rhymes than in non-rhymes (53% vs. 47%), which contradicts the result of the current study. As a comparatively unproductive plural allomorph, *-e* with umlaut occurs less frequently with non-rhymes than *-e* without umlaut and thus it would reduce the proportion of the allomorph "*-e* with or without umlaut" used with non-rhymes.

Different calculation methods and different constellations of independent variables can also influence the distribution of significance values. In the binary logistic regressions calculated within GLMM – a more conservative method than categorical regressions – only few factors delivered significant results. Whereas the results for *s*-forms turned out to be fairly stable and did not depend much on the calculation method, *en*-forms yielded between two and five *p*-values below .05. It is, again, the task of the researcher to decide which model should be reported. Due to limitations of space, fixed coefficients were not given in the tables. Fortunately, there were no contradictions between fixed effects and coefficients in terms of positive or negative associations. These corresponded to the tendencies described in the cross-tables. In linear regressions (Table 2), however, not a single significant fixed coefficient was identified because the “broad” data design tends to deliver less specified results in comparison with the “long” design.

Further factors that would have influenced the significance of results in GLMM are, among others, interactions between variables (e.g. “word-final phoneme”*“grammatical gender”), number of iterations, use of model-based or robust covariances, and order of inclusion of independent variables. Although the latter should be irrelevant (or, at least, its relevance is not described in the *SPSS* documentation or anywhere else to our knowledge), placement of variables at the beginning of the list slightly increases the probability of a significant result in comparison with the placement at the end of the list. Numerous other variables might be relevant for the distribution of plural markers and could be included in the regressions: sociolinguistic and demographic characteristics of test subjects (e.g. educational level, immigration and dialectal background), characteristics of plural markers (e.g. iconicity, frequency), and also some further characteristics of test items (e.g. relatively rare phoneme combinations /bn/, /fn/, /sn/; gender variation within the real nouns such as sg. *der Bund* ‘alliance’ → pl. *Bünde*, sg. *das Bund* ‘bundle’ → pl. *Bunde*; a phonological/orthographic overlap between more frequent and more phonologically close nouns, cf. nonce word *Klot*: more frequent *Brot* ‘bread’ vs. more phonologically close, but less frequent *Schlot* ‘chimney’).

The tests used demonstrated that certain differences in the distribution of plural markers depending on the classification of nouns as rhymes or non-rhymes did exist, although this variable played a minor role (or no statistically significant role at all according to GLMM) in comparison with grammatical gender (including gender shift), word-final phonemes, associations with existing German words, and the presence of monophthongs or diphthongs that can be subject to umlauting.

No evidence was found that *-s* was used as the default plural marker in non-rhyming nonce words and that other plural markers can be considered to be irregular. In fact, *-s* was used only in 16% of plural formations with non-rhymes, which was 8% more than with rhymes, but still arguably not enough to speak of a dominant role. Apart from *-s*, in non-rhymes compared to rhymes, a significantly higher frequency of *-(e)n* and *-e* was identified: *-(e)n* also made up 8% more of all pluralizations with non-rhymes than with rhymes, and *-(e)*, in spite of the difference of only 2% between non-rhymes and rhymes, accounted for almost half of all pluralizations with both kinds of nonce words. This is not just a quantitative, but a qualitative difference from the results of Marcus et al. (1995), which was, however, to be expected because systematic mismatches between acceptability ratings and production frequencies have been reported in previous research (Kempen & Harbusch 2005; Arppe & Järvikivi 2007; Bader & Häussler 2010).

All three markers (*-s*, *-(e)n*, *-e*) are iconic, productive, and the latter two are also the most frequent ones in Modern High German (Zaretsky et al. 2011). All three markers do not require umlauting, which allows to avoid (potentially wrong) modifications of the stems of unknown words. The plural allomorph *-s*, even though infrequent, is phonotactically highly compatible and semantically associated with any unusual language material, which might have increased its frequency in non-rhymes. Obviously, with language material like the nonce words chosen by Marcus et al. (1995), that is, nouns with very few cues on possible plural forms, there is no considerable difference between pluralization schemata of children and adults, because children are known to prefer the same three plural markers with nonce words (Zaretsky et al. 2013c). Other plural markers (*-er*, umlaut, *-e* with umlaut) are hardly productive in German and, therefore, occur comparatively rarely in the pluralizations of both children and adults.

Differences between the results presented here and those reported in Marcus et al. (1995) can be traced back to two factors: first, a very limited sample size and some other more or less problematic issues in the statistical analysis by Marcus et al. (1995); second, plausibility tasks in Marcus et al. (1995) versus active plural production in the study presented here, the latter explanation probably being more relevant. Plausible forms are not necessarily the forms preferred by test subjects. Therefore, active pluralization might deliver not only quantitatively, but also qualitatively different results as regards internalized pluralization rules and strategies. The results presented demonstrate that plural forms produced by German adults can most adequately be explained in terms of single-route models, without subdivision of plural markers into default and irregular ones.

References

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159.

Baayen, Harald R. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bader, Markus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2). 273–330.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form.* Amsterdam: Benjamins.

Clahsen, Harald. 1999. Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22(6). 991–1060.

Clahsen, Harald, Monika Rothweiler, Andreas Woest & Gary F. Marcus. 1992. Regular and irregular inflection in the acquisition of German noun plurals. *Cognition* 45(3). 225–255.

Dressler, Wolfgang U., Willi Mayerthaler, Oswald Panagl & Wolfgang U. Wurzel. 1987. *Leitmotifs in natural morphology*. Amsterdam: Benjamins.

Elsen, Hilke. 2001. The acquisition of German plurals. In Sabrina Bendjaballah, Wolfgang U. Dressler, Oskar E. Pfeiffer & Maria D. Voeikova (eds.), *Morphology 2000: Selected papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000,* 117–128. Amsterdam: Benjamins.

Fakhry, Salah A. 2005. *Die Entwicklung des deutschen Pluralsystems im 20. Jahrhundert*. Marburg: Philipps-University Marburg dissertation.

Faul, Franz, Edgar Erdfelder, Axel Buchner & Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41(4). 1149–1160.

Hahn, Ulrike & Ramin C. Nakisa, 2000. German inflection: Single route or dual route? *Cognitive Psychology* 41(4). 313–360.

IBM. 2012. *IBM SPSS Statistics*. Armonk: IBM Corp.

Institut für Deutsche Sprache. 2009. Korpusbasierte Wortformenliste DeReWo, v-100000t-2009-04-30-0.1, mit Benutzerdokumentation. Mannheim: Institut für Deutsche Sprache, http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html (November 20, 2015.)

Kempen, Gerard & Karin Harbusch. 2005. The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, 329–349. Berlin: Mouton de Gruyter.

Köpcke, Klaus-Michael. 1993. *Schemata bei der Pluralbildung im Deutschen*. Tübingen: Narr.

Korecky-Kröll, Katharina & Wolfgang U. Dressler. 2009. The acquisition of number and case in Austrian German nouns. In Ursula Stephany & Maria D. Voeikova (eds.), *Development of nominal inflection in first language acquisition: A cross-linguistic perspective*, 265–302. Berlin: Mouton de Gruyter.

Korecky-Kröll, Katharina, Gary Libben, Nicole Stempfer, Julia Wiesinger, Eva Reinisch, Johannes Bertl & Wolfgang U. Dressler. 2012. Helping a crocodile to learn German plurals: Children's online judgment of actual, potential and illegal plural forms. *Morphology* 22(1). 35–65.

Marcus, Gary F., Ursula Brinkmann, Harald Clahsen, Richard Wiese & Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive Psychology* 29(3). 189–256.

Mugdan, Joachim. 1977. *Flexionsmorphologie und Psycholinguistik*. Tübingen: Narr.

Niedeggen-Bartke, Susanne. 1999. Flexion und Wortbildung im Spracherwerb. In Jörg Meibauer & Monika Rothweiler (eds.), *Das Lexikon im Spracherwerb*, 208–228. Tübingen: Francke.

Ruoff, Arno. 1981. *Häufigkeitswörterbuch gesprochener Sprache*. Tübingen: Niemeyer.

Schüller, Katharina. 2015. *Statistik und Intuition*. Heidelberg: Springer.

Spreng, Bettina. 2004. Error patterns in the acquisition of German plural morphology: Evidence for the relevance of grammatical gender as a cue. *Toronto Working Papers in Linguistics* 23(2). 147–172.

Szagun, Gisela. 2001. Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language* 21(62). 109–141.

Tutz, Gerhard. 2011. *Regression for categorical data*. Cambridge: Cambridge University Press.

Wegener, Heide. 1994. Variation in the acquisition of German plural morphology by second language learners. In Rosemarie Tracy & Elsa Lattey (eds.), *How tolerant is Universal Grammar? Essays on language learnability and language variation*, 267–294. Tübingen: Niemeyer.

Wurzel, Wolfgang U. 1984. *Flexionsmorphologie und Natürlichkeit*. Berlin: Akademie-Verlag.

Zaretsky, Eugen & Benjamin P. Lange. 2014. Influence of intra- and extralinguistic factors on the distribution of plural allomorphs in German. *California Linguistic Notes* 39(1). 73–114.

Zaretsky, Eugen & Benjamin P. Lange. 2015. Über die trügerische Sicherheit der statistischen Signifikanz in der klinischen Linguistik, am Beispiel des Zusammenhangs zwischen Dysphonie und Sprachleistungen deutscher Vorschulkinder. Presentation at the conference "Bundesverband Klinische Linguistik. Workshop", Mainz, Germany, 30 April–02 May 2015.

Zaretsky, Eugen, Benjamin P. Lange, Harald A. Euler & Katrin Neumann. 2011. Pizzas, Pizzen, Pizze: Frequency, iconicity, cue validity, and productivity in the plural acquisition of German preschoolers. *Acta Linguistica* 5(2). 22–35.

Zaretsky, Eugen, Benjamin P. Lange, Harald A. Euler & Katrin Neumann. 2013a. Differences in plural forms of monolingual German preschoolers and adults. *Lingue e Linguaggi* 10. 169–180.

Zaretsky, Eugen, Benjamin P. Lange, Harald A. Euler & Katrin Neumann. 2013b. Acquisition of German pluralization rules in monolingual and multilingual children. *Studies in Second Language Learning and Teaching* 3(4). 551–580.

Zaretsky, Eugen, Katrin Neumann, Harald A. Euler & Benjamin P. Lange. 2013c. Pluralerwerb im Deutschen bei russisch- und türkischsprachigen Kindern im Vergleich mit anderen Migranten und monolingualen Muttersprachlern. *Zeitschrift für Slawistik* 58(1). 43–71.

Appendix

Non-rhymes: *ein(e) Bnaupf, ein(e) Pläk, ein(e) Plaupf, ein(e) Snauk, ein(e) Bneik, ein(e) Pleik, ein(e) Fnöhk, ein(e) Bröhk, ein(e) Pröng, ein(e) Fnähf, ein(e) Pnähf, ein(e) Fneik*

Table. Rhymes with the most probable associations based on frequency lists of types (Ruoff 1981) and tokens (Institut für Deutsche Sprache 2009)

	High-frequency rhymes among existing German nouns	
Item	**Rank 1**	**Rank 2**
Pisch		
Type	*der Tisch* ('table') – *Tische*	*der Fisch* ('fish') – *Fische*
Token	*der Tisch* ('table') – *Tische*	*der Fisch* ('fish') – *Fische*
Bral		
Type	*das Tal* ('dale') – *Täler*	*das Mal* ('time, mark') – *Male*
Token	*das Mal* ('time, mark') – *Male*	*der Saal* ('hall') – *Säle*
Pind		
Type	*das Kind* ('child') – *Kinder*	*der Wind* ('wind') – *Winde*
Token	*das Kind* ('child') – *Kinder*	*der Wind* ('wind') – *Winde*
Kach		
Type	*das Dach* ('roof') – *Dächer*	*der Bach* ('stream') – *Bäche*
Token	*das Dach* ('roof') – *Dächer*	*der Bach* ('stream') – *Bäche*
Pund		
Type	*der Grund* ('reason, ground') – *Gründe*	*das Pfund* ('pound') – *Pfunde*
Token	*der Grund* ('reason, ground')– *Gründe*	*der Bund* ('alliance') – *Bünde*
Klot		
Type	*das Brot* ('(loaf of) bread') – *Brote*	*die Not* ('need') – *Nöte*
Token	*die Not* ('need') – *Nöte*	*das Boot* ('ship') – *Boote*
Vag		
Type	*der Tag* ('day') – *Tage*	*der Schlag* ('strike') – *Schläge*
Token	*der Tag* ('day') – *Tage*	*der Schlag* ('strike') – *Schläge*
Spert		
Type	*der Wert* ('value') – *Werte*	*das Pferd* ('horse') – *Pferde*
Token	*der Wert* ('value') – *Werte*	*das Pferd* ('horse') – *Pferde*
Mur		
Type	*die Uhr* ('watch') – *Uhren*	*die Schnur* ('cord') – *Schnüre*
Token	*die Uhr* ('watch') – *Uhren*	*die Spur* ('trace') – *Spuren*
Raun		
Type	*der Zaun* ('fence') – *Zäune*	—
Token	*der Zaun* ('fence') – *Zäune*	—
Nuhl		
Type	*der Stuhl* ('chair') – *Stühle*	—
Token	*der Stuhl* ('chair') – *Stühle*	—
Spand		
Type	*die Hand* ('hand') – *Hände*	*das Land* ('country') – *Länder*
Token	*das Land* ('country') – *Länder*	*die Hand* ('hand') – *Hände*

University
of Bamberg
Press

Over the past few decades, linguistic theorizing has benefited from an increasing trend towards empirical methodologies across all disciplines. Methodological know-how – both productive and receptive – has thus become one of the key qualifications for researchers. The empirical turn in linguistics has gone hand in hand with a considerable diversification of research methods. This diversity, which has come to be seen as a strength of linguistics as a field, has also benefited linguistic theory building. The present volume contains selected contributions from the 2015 Methods and Linguistic Theories (MaLT) symposium that address the aforementioned issues from an empirical and/or theoretical perspective. They can be seen as the essence of what MaLT was about, and illustrate the range of topics covered as well as the various concerns and approaches that featured during the event.

eISBN: 978-3-86309-458-4

9 783863 094584

www.uni-bamberg.de/ubp