Quantitative Methods in the Humanities
and Social Sciences

Masako Fidler
Václav Cvrček *Editors*

Taming the Corpus

From Inflection and Lexis to Interpretation

Springer

Quantitative Methods in the Humanities and Social Sciences

Editorial Board
Thomas DeFanti, Anthony Grafton, Thomas E. Levy, Lev Manovich,
Alyn Rockwood

Quantitative Methods in the Humanities and Social Sciences is a book series designed to foster research-based conversation with all parts of the university campus – from buildings of ivy-covered stone to technologically savvy walls of glass. Scholarship from international researchers and the esteemed editorial board represents the far-reaching applications of computational analysis, statistical models, computer-based programs, and other quantitative methods. Methods are integrated in a dialogue that is sensitive to the broader context of humanistic study and social science research. Scholars, including among others historians, archaeologists, new media specialists, classicists and linguists, promote this interdisciplinary approach. These texts teach new methodological approaches for contemporary research. Each volume exposes readers to a particular research method. Researchers and students then benefit from exposure to subtleties of the larger project or corpus of work in which the quantitative methods come to fruition.

More information about this series at http://www.springer.com/series/11748

Masako Fidler • Václav Cvrček
Editors

Taming the Corpus

From Inflection and Lexis to Interpretation

Springer

Editors
Masako Fidler
Department of Slavic Studies
Brown University
Providence, RI, USA

Václav Cvrček
Institute of the Czech National Corpus
Charles University
Prague 1, Czech Republic

ISSN 2199-0956 ISSN 2199-0964 (electronic)
Quantitative Methods in the Humanities and Social Sciences
ISBN 978-3-319-98016-4 ISBN 978-3-319-98017-1 (eBook)
https://doi.org/10.1007/978-3-319-98017-1

Library of Congress Control Number: 2018957139

© Springer Nature Switzerland AG 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction

Václav Cvrček and Masako Fidler

Empirical linguistics has always gravitated towards quantification. With the advent of electronic corpora—large, searchable sets of natural language data, quantification has become part and parcel of linguistic studies. In the past few decades in particular, we have witnessed a "quantitative turn" in various schools of linguistics (cf. Janda, 2013 for cognitive linguistics) and in the digital humanities which was further accelerated by the advent of text corpora. This volume aims to showcase a variety of recent quantitative approaches that "tame the corpus"; it shows how language corpora can be used for research questions of interest to students and scholars in the humanities and social scientists.[1] It simultaneously fills a lacuna in mainstream English-based quantitative linguistic studies by demonstrating that quantitative methods applied on inflectional language may reveal novel phenomena.

This introduction presents our position with respect to quantitative language data analysis. We first revisit the apparent "quantitative–qualitative dichotomy" to show that there are features shared by quantitative and qualitative analyses. We then discuss the advantages of quantitative data and statistical evaluation. The chapter closes with an overview of the studies in this volume.

[1] The volume was inspired by the Workshop on Quantitative Text Analysis for the Humanities and Social Sciences, which the editors organized at Brown University on April 8 and 9, 2016.

V. Cvrček (✉)
Institute of the Czech National Corpus, Charles University, Prague 1, Czech Republic
e-mail: vaclav.cvrcek@ff.cuni.cz

M. Fidler
Department of Slavic Studies, Brown University, Providence, RI, USA
e-mail: masako_fidler@brown.edu

© Springer Nature Switzerland AG 2018

M. Fidler, V. Cvrček (eds.), *Taming the Corpus*, Quantitative Methods in the Humanities and Social Sciences, https://doi.org/10.1007/978-3-319-98017-1_1

A Quantitative–Qualitative Dichotomy

Quantitative and qualitative approaches are commonly viewed in opposition to each other. The comparison between the two approaches potentially leads to oversimplification[2]: quantitative approaches are often considered more reliable, more precise, more inductive, and allow more convincing generalizations and hypothesis testing than qualitative approaches; qualitative approaches are viewed as subjective, focused on a specific instance, exploratory (allowing for defining the problem or establishing a hypothesis), and deductive. Contrary to such a popular impression, however, each approach has its strengths and weaknesses. Qualitative research may obtain in-depth knowledge of a particular sample (e.g., through the close reading of a single literary text), revealing a wide range of questions/hypotheses about the text (e.g., metaphors used, prominent motifs, intertextual links, and allusions). The trade-off is that the researcher's claim is based on a small sample. Quantitative research usually starts with a narrowly focused observation (e.g., the relative prominence of individual words) from a larger population (e.g., the entire corpus of texts written by one author or texts of one epoch). This type of research may lead to overarching conclusions. Its trade-off is that many details may be omitted as unimportant or irrelevant to the research question. In other words, we may either examine a small number of instances of the phenomenon under scrutiny very carefully, or a large number of instances superficially. Regardless of efforts and funds, each type of research has its own omnipresent trade-off.

Quantitative and qualitative approaches moreover share certain properties. Qualitative research may involve some minimum "quantification" when some recurrent patterns are noted.[3] Quantitative research presupposes a "qualitative delimitation" of categories: for example, types of nouns or parts of speech must be qualitatively defined before their frequencies can be calculated. To cite Herdan, "[t] here is no sharp dividing line between qualitative and quantitative methods, but only transition comparable to that from large scale to small sca[l]e maps" (1966, p. 2). If neither approach can exist in isolation, then we can expect that both approaches would also *share* some advantages as well as disadvantages.

One crucial concept to capture such advantages and disadvantages of both approaches is *reductionism*. In any research—qualitative and quantitative alike, we have to make a decision on what to include in our investigation. Researchers usually pick only those available (or noticeable) features that appear relevant to the research question and ignore the rest. Consequently, each description is shaped by a

[2] Superficial Internet search often leads one to have such an impression, cf. https://www.orau.gov/cdcynergy/soc2web/content/phase05/phase05_step03_deeper_qualitative_and_quantitative.htm and https://keydifferences.com/difference-between-qualitative-and-quantitative-research.html#ComparisonChart. Accessed 25 May 2018.

[3] Even a singular appearance represents quantity (=1) and the difference between a single or no occurrence may result in ascribing an important property to the phenomenon under examination or not. But usually, even in qualitative studies, multiple examples demonstrating a hypothesis are better than one.

combination of what has been found and what has been left aside (either knowingly or unknowingly): we select specific categories, terms, a point of view, and/or a methodology. This problem of reducing the research input is usually mentioned in relation to quantitative studies; in order to examine some phenomenon quantitatively, we have to zoom in on a limited and manageable amount of features. But the same problem can be found in qualitative research as well; the researcher may consider a broader context of relations interacting with the target phenomenon, but it is impossible to include all the potential influences (e.g., all intertextual links). What usually happens in qualitative analysis is that the researcher discusses only those aspects of his/her choice, to the exclusion of other aspects.[4] Both quantitative and qualitative approaches thus may suffer from reductionism to varying degrees.

Likewise, a degree of *reliability* is of concern to both quantitative and qualitative studies. It is likely that examination of a large sample (at the corpus level) leads to substantive conclusions about the target language phenomenon. The reliability of the researcher's findings, however, will depend on the level of reductionism: reducing a complex system to a few easy-to-quantify variables may point to interesting results, but this inevitably leads to a schematic description with some important parts missing. On the other hand, if one examines the same research question qualitatively in a single text with an eye to a wide range of interacting factors, the study may yield valid results so long as its findings can be applied to other texts. In order to achieve reliable results, then, we need both methods.

Degrees of reductionism can also affect degrees of *objectivity* and *subjectivity*—properties that are often attributed to quantitative and qualitative research, respectively. Quantitative methods can be qualified as objective, *provided* that the categories they use (e.g., parts of speech, as mentioned above) are validated by convincing qualitative research.

There is yet another property that supposedly divides qualitative and quantitative methods: *inductive* vs. *deductive reasoning*. Qualitative methods are often associated with the former and quantitative methods with the latter (Rasinger, 2008, p. 11). In quantitative studies, it is common practice to impose a statistical model on the data (especially in situations where many models are available) based on our general assumptions about the gathered evidence; this approach clearly involves deductive reasoning. However, we may find also counterexamples. Corpus-driven (Tognini-Bonelli, 2001) or data-driven quantitative studies are built on inductive reasoning; they assume that the theory has to be optimized for large amounts of data (and not the other way around). As for qualitative studies, often described as inductive, they can be deductive by approaching the target subject with pre-formulated theory or by describing the subject within an established concept or point of view (as in critical discourse analysis). Clearly, the boundaries between quantitative and qualitative studies are not as discrete as they appear.

[4] Unlike many quantitative studies, where the amount of reduction is sometimes explicitly acknowledged. Johnson states that in fact any (statistical) inference about the data is guessing; what quantitative methods can help us with is to quantify how reliable our guesses are (2008, p. 3).

Furthermore, there is also a perception that qualitative study yields a *hypothesis*, which should consequently be *tested* quantitatively. This is not always the case. Both qualitative and quantitative approaches share an exploratory potential. Sometimes, the underlying phenomena are visible only from the perspective of larger data (collocations in corpus linguistics being an obvious example). Sometimes, important aspects can be spotted only through detailed qualitative study. New hypotheses may arise from both directions.

Why the Use of Corpus and Quantitative Methods?

In spite of the shared features between qualitative and quantitative methods, the latter nonetheless has significant additional and possibly more important advantages, given the increasing need for empirical evidence in linguistics. One of them—as we as editors see it—is that quantitative methods are likely to produce testable (or falsifiable, cf. Popper, 1959 [2005]) outcomes. There are two important aspects of quantitative methods: each result can be replicated on the original data (everyone is allowed to rerun the experiment and verify if the reported results are based on solid analysis); and each method can be normally applied to different data (which allows for testing the limits of generalization). In contrast, qualitative analysts lacking large data sets and statistics would have to make extraneous efforts to do the same.

The second advantage of a quantitative approach is that it is supported by existing mathematical and statistical methods. An elaborated system of dealing with quantifiable variables already exists ready to use, with well-described (although sometimes complex and hard to understand) limitations and pitfalls. In addition, mathematics is an artificial system that does not bear any false connotations. In order to understand why this is an advantage, we must recognize that there is a metaphor at the core of any scientific description (e.g., the development of languages as a tree spreading out branches). By translating language features into counts and frequencies, we use a mathematical "metaphor," which has the advantage of being a universally comprehensible but simultaneously artificial system unburdened by connotations. This property is hard to find outside of mathematics.

The third advantage of quantitative approaches is that they allow "*interobjectivity*"—the possibility of seeing similar patterns in different fields of study. By this principle, we may compare such things as the similarity of word frequency distribution (known as Zipfian distribution) to the distribution of population within the cities of a country. By recognizing similar patterns across different disciplines and objects of study, we can enhance our own understanding of language and bring new inspiring ideas into its description.

Finally, there is a practical motivation to use quantitative methods. Although both quantitative and qualitative studies may be empirical, only the former assumes that generalization is possible only after the examination of representative data samples. This was not an issue in the past, but with the advent of large electronic corpora, one now has to search for a method capable of *taming the once unthinkable amount of data*.

Taming the Corpus

Quantification, with all its shortcomings and deficiencies, is still the only way to deal with the large corpora, which are increasingly used to produce findings about language, literature, and society. Besides describing linguistic phenomena, such as collocability of words (e.g., Gries, 2013) or language variability (e.g., Biber and Conrad, 2009) to name at least a few, quantitative methods applied to large language data empower scholars to explore social issues, e.g., media portrayal of refugees and asylum seekers (Baker and McEnery, 2005). Quantitative methods also help capture global themes predominant in the national literatures and historical documents (Jockers, 2013).

Such studies largely focus on the lexicon, which plays several important roles in the production of text and our perception of the world. Words occurring at unexpectedly high frequencies, for example, point to prominent topics—word frequencies can reveal what readers find striking in a text, especially when contrasted against a background of other corpora. Word clusters can help identify phrases or formulaic expressions in large collections of discourse samples. The use of such lexicon-centered methods understandably originated from the study of texts in English, a language with little explicit grammatical marking.

This book examines lexis as well as smaller grammatical units that can be objectively identified—detailed components in phonology and morphosyntax (syllable structure, modifier-modified agreement, and grammatical case). This line of research is made possible by the explicit grammatical marking of Czech and the large and well-documented language data available through the Czech National Corpus (henceforth CNC). CNC (see https://www.korpus.cz) is one of the most robust and well-balanced language corpora in the world and the most developed corpus of any Slavic language. Since its establishment in 1994, the CNC project has been continuously mapping Czech in different domains; several series of corpora have been developed and maintained, namely a synchronic written corpus (currently with four billion words), a spoken corpus (focusing on unprepared informal dialogues with 6.4 million words), and a diachronic corpus (covering the period from the fourteenth century to 1945). CNC also contains parallel-language corpora (InterCorp) that facilitate contrastive research in more than thirty languages (245 million words in Czech, 1.87 billion in aligned texts of other languages); InterCorp is valuable not only for its size (it is one of the largest and the most diverse among the Slavic parallel corpora available) but also for its careful design and manually checked core section in fiction. Moreover, CNC is equipped with web-based software tools with continually updated functions. These tools ensure a large number of possibilities to probe language on multiple levels: translation between languages, collective perceptions of language, and analysis of literary and political texts.

The aim of this book is to showcase multiple approaches to language, literature, and society. The volume demonstrates diverse methods, which range from "simple" quantification as a means of description to sophisticated statistical methods employed for the purpose of revealing new phenomena.

Section 1 (*Words, rhymes, and grammatical forms*) deals with phonotactics, poetic structure, morphological complexity used to differentiate literary style, and native speakers' sense of grammaticality—issues pertinent to linguistic typology, cognition and language, and literary studies. The article by Neil Bermel, Luděk Knittl, and Jean Russel probes the relationship between language exposure and speakers' performance on production and ratings tasks. Frequency data from CNC is used as a proxy for language exposure. Jiří Milička and Hana Kalábová explore vowel phonotactics in Czech words and word stems. The authors identify s vowel length and vowel front-/backness. Radek Čech and Miroslav Kubát propose a computational method to measure the morphological richness of texts (an index of utmost importance in inflected languages), thereby finding a way to quantitatively characterize author styles. Petr Plecháč applies a quantitative method to poetry. The author develops a method to identify frequent rhyme pairs in poetry corpus by collocation extraction technique and uses the output as a training set for machine learning. The method is tested on poetry corpora in three different languages (Czech, English, and French) with high accuracy.

Section 2 (*Not only "lost" in translation*) takes us to interlanguage relations. Lucie Chlumská takes the "top-down view." She compares the prominent n-grams and POS-grams (n-grams consisting of part-of-speech tags) in translated Czech and in the English source texts. She examines the viability of "translation universals" that are independent of linguistic similarities or differences between the original and the translated texts. While confirming such universal tendencies in Czech–English translations, the author argues that no component claimed to belong to the category of a translation universal can be distinctly isolated; translated texts manifest a combination of properties. Moreover, the author discusses the specificities of cross-linguistic comparison based on POS-grams and n-grams in the two typologically different languages. David Danaher takes the bottom-up view, looking at the specific sociocultural contexts in which lexis is embedded. He analyzes collocations to study the semantics of *lidskost* (often translated as "humanity," "humanness," or "humaneness") and related words as used in Václav Havel's writings. Combining quantitative and qualitative methods, the author traces the contexts that molded the semantics of these words. Danaher's collocation analysis illustrates how words come to defy translation because of their usage in socioculturally specific contexts that have evolved over the past centuries. The issues in this section are important regardless of the size of the target language (the language into which a text is translated). Admittedly, complexity in translation is an issue for midsized and smaller languages as target languages since translated texts constitute a large part of literary production. However, it is also an important issue in larger target languages spoken by a large monolingual population that has little access to the original texts.

Section 3 (*Understanding discourse*) demonstrates how quantitative analysis of texts can contribute to our understanding of society and connects the volume to legal language (Kieran Williams), construction of gender (Adrian Zasina), and discourse position and implicit ideology (Masako Fidler and Václav Cvrček). Williams' study demonstrates how collocations can identify potential costs of the general public's misunderstanding legal language. As an illustration, the author uses words from the 2017 Czech gun bill, written with the intention of creating a constitutional right to keep and bear arms, to assist the state in protecting national security. By comparing the usage of crucial terms used both in the gun law and in non-legal texts, Williams suggests a “marked misalignment” between the two usages that could gravely affect compliance with and enforcement of the gun law. Zasina uses corpus data to investigate gender representation of politicians in Czech daily newspapers. His study serves as a springboard to consider a need to go beyond identifying explicit gender stereotypes, and to construct a more complex conceptual model to interpret subtle attributes used on male and (especially) female politicians. Fidler and Cvrček take the basic concept of keyword analysis, a corpus linguistic method used to identify prominent words in text (“aboutness”), as a starting point, but both extend and add to its functionality. The “Multi-level Discourse Prominence Analysis” provides information about a text’s overarching rhetoric and helps to objectivize the ideological content of news. It takes advantage of the inflectional morphology of Czech (via analysis of prominent morphs) to unpack implicit and recurrent messages in texts, and more importantly has the potential to reveal implicit ideology at a deeper (perhaps subconscious) level.

Taming the Corpus presents a variety of quantitative approaches to language, literature, and society. The volume attempts to show how quantitative methods can be further empowered by utilizing features that are characteristic of an inflectional language. The editors hope that the book will spark interest in thus-far underutilized grammatical markings in many other languages that could potentially enhance objectivity and precision in quantitative methods.

Acknowledgments The publication of this volume was made possible by support from grant *Progres* Q08 *Czech National Corpus* implemented at the Faculty of Arts, Charles University and the Humanities Research Grant from Brown University. Special thanks goes to Mathew Amboy and Faith Su from Springer who saw through the entire publication process and Marek Nekula for thoughtful and helpful comments on the manuscripts. The editors would also like to thank Andrew Malcovsky for copyediting work. Last but not least, many thanks to Lída Cvrčková Porkertová and Vlastimil Fidler for their support and patience.

References

Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics, 4*(2), 197–226.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics, 18*(1), 137–165.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Berlin, Germany: Springer.

Janda, L. A. (Ed.). (2013). *Cognitive linguistics: The quantitative turn*. Berlin, Germany: De Gruyter Mouton.

Jockers, M. L. (2013). *Macroanalysis. Digital methods and literary history*. Urbana, IL: University of Illinois Press.

Johnson, K. (2008). *Quantitative methods in linguistics*. Malden, MA: Blackwell publishing.

Popper, K. (1959) [2005]. *The logic of scientific discovery*. London, UK: Routledge.

Rasinger, S. M. (2008). *Quantitative research in linguistics. An introduction*. London, England: Continuum.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Part I
Words, Rhymes, and Grammatical Forms

Chapter 2
Do Users' Reading Skills and Difficulty Ratings for Texts Affect Choices and Evaluations?

Neil Bermel, Luděk Knittl, and Jean Russell

Abstract In our contribution, we consider how corpus data can be used as a proxy for the written language environment around us in constructing offline studies of native-speaker intuition and usage. We assume a broadly emergent perspective on language: in other words, the linguistic competence of individuals is not identical or hardwired but forms gradually through exposure and coalescence of patterns of production and reaction. We hypothesize that while users presumably all in theory have access to the same linguistic material, their actual exposure to it and their ability to interpret it may differ, which will result in differing judgments and choices. Our study looks at the interaction between corpus frequency and two possible indicators of individual difference: attitude towards reading tasks and performance on reading tasks. We find a small but consistent effect of task performance on respondents' judgments but do not confirm any effects on respondents' production tasks.

Keywords Czech morphology · Variation · Overabundance · Acceptability judgments · Experimental linguistics · Usage-based approach

Introduction[1]

Considerable attention has been devoted to whether all native speakers of a language access the same linguistic structures and material in similar ways, and whether, having accessed it, their use of and reaction to language (what we will call *linguistic behavior*) differ as well in predictable ways. There is accumulating

[1] This research was carried out as part of the project "Acceptability and forced-choice judgements in the study of linguistic variation," funded by the Leverhulme Trust (RPG-407). The support of the Trust is gratefully acknowledged.

N. Bermel (✉) · L. Knittl · J. Russell
University of Sheffield, Sheffield, UK
e-mail: n.bermel@sheffield.ac.uk; l.knittl@sheffield.ac.uk; j.russell@sheffield.ac.uk

© Springer Nature Switzerland AG 2018

M. Fidler, V. Cvrček (eds.), *Taming the Corpus*, Quantitative Methods in the Humanities and Social Sciences, https://doi.org/10.1007/978-3-319-98017-1_2

evidence that intra-speaker variation can point to differences in linguistic behavior that are not random or insignificant.

We can propose that speakers' varying backgrounds (i.e, their *exposure* to language) affect language in use (i.e, their *output* or their *evaluation of input*). In other words, if we call what underlies this linguistic behavior a "grammar," each speaker's is subtly different. Corpus data can, if carefully used, be hypothesized to represent this "exposure" to at least the written form of the language, which is the tack we will take in this study.[2] In doing so, we aim to add to the evidence showing how corpus frequency can be useful in detecting and predicting our use of language.

Background

Evidence has, at times, pointed to vocabulary size, education, profession, and reading recall abilities as factors differing from subject to subject that affect one's "personal" linguistic behavior, and these differences have been found in syntax, word-formation, and inflectional morphology. While we might try to explain away differences resulting from regional or age variation as the product of language shift and change, it is harder to do so with e.g. educational or professional differences.

In a series of articles, Dąbrowska has tracked some of these differences in speaker backgrounds, which, she shows, lead to differences in both linguistic performance and linguistic judgments. Dąbrowska (2008) looked at a sample of users stratified by educational background and assessed their performance on a production task. She concluded that "the results… revealed large individual differences in speakers' ability to inflect unfamiliar nouns which were strongly correlated with education" (2008, p. 941). Having attempted to eliminate some possible confounding factors, she concluded, "We can be reasonably confident… that the observed differences in scores in the other conditions reflect genuine differences in linguistic proficiency" (2008, p. 945). A logical deduction from that might have been that more educated speakers had larger vocabularies; however, Dąbrowska did not find enough evidence for this, saying, "…the results do not support the hypothesis that the critical variable is vocabulary size, although they do not unequivocally rule it out" (2008, p. 949). In a later study, she examined judgments of sentence well-formedness given by linguists and nonlinguists, and found that:

> Linguists' judgments are shown to diverge from those of nonlinguists. These differences could be due to theoretical commitments (the conviction that linguistic processes apply 'across the board,' and hence all sentences with the same syntactic structure should be equally grammatical) or to differences in exposure (the constructed examples of this structure found in the syntactic literature are very unrepresentative of ordinary usage) (2010, p. 1).

[2] Fidler and Cvrček's (2015) study of keyword analysis in Czech presidential New Year speeches uses this approach to good effect to demonstrate how different types of exposure, in the guise of reference corpora, can be used to model differing potential receptions of a text.

While Dąbrowska was cautious in her conclusions about whether educational differences and vocabulary size can be so closely linked, other researchers have made the connection between linguistic behavior and vocabulary size more directly. For example, Frisch & Brea-Spahn (2010) found that vocabulary size, as measured by the results of a word familiarity rating task, correlates with acceptability scores on a word-formation task. They noted:

> Participants with a larger vocabulary in English were more accepting of low probability nonwords in English. It appears that those with greater vocabulary knowledge are more likely to have experienced improbable phonological constituents, and may also have a lower threshold for "unacceptable" nonwords, if their threshold is based on a likelihood estimate from their individual lexicon (2010, p. 345).

Reading abilities also affect judgments: Staum Casasanto, Hofmeister, & Sag (2010) investigated how differences in reading span interact with judgements.[3] Reading span task scores were highly significant predictors of acceptability scores on a task involving the syntax of embedded clauses, e.g, *The nurse from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room* (Staum Casasanto et al. 2010, p. 224). They concluded that

> [P]articipants' reading span scores predict sentence judgments differently for different types of manipulations. Participants with higher reading spans tend to judge ungrammatical sentences as being worse than their low-span counterparts do, yet they tend to judge difficult sentences as being better than participants with lower reading spans (2010, p. 228).

A further set of factors that have been shown to contribute to analyses of linguistic behavior are those that derive from analyses of the task performance itself. For example, Divjak demonstrates that ratings given on "filler" items—in other words, items designed to distract the respondent, rather than the test items themselves—are in fact the best predictor of how a respondent rates the test items (in this instance manipulating the complement of certain verbs). This suggests that an overall *individual* variation in how people use rating scales can account for some of the differences we see; Divjak terms this "non-linguistic variability" (2016 [2017], p. 14). Bermel, Knittl, & Russell show that respondents' ratings of the *less* common of two variants are the best predictor of how they answer on a production task. In other words, looking at the ratings for the lesser-used ending {a} in the genitive singular rather than the more-common {u} gives us the best chance of predicting which ending native speakers will insert in a gap-filling task (2015a, pp. 304–306).

In summary, then, it seems that a variety of speaker-specific factors can influence linguistic behavior. Some of these, such as educational attainment and profession,

[3] Reading span tasks ask participants to read unconnected sentences, memorizing the final word of each sentence, which they then must recall later. There is some dispute about what exactly they are measuring (Hupet, Desmette, & Schelstraete, 1997), but as Conway et al. point out, they have been widely used nonetheless to assess how we tap into our working memory's storage and processing functions: "The task is essentially a simple word span task, with the added component of the comprehending of sentences. Subjects read sentences and, in some cases, verify the logical accuracy of the sentences, while trying to remember words, one for each sentence presented" (Conway et al., 2005, p. 771).

appear to be nonlinguistic factors but may in fact be linked to an individual's linguistic abilities. Others, including vocabulary size (either measured via the self-reported familiarity of words or accuracy on a semantics test) and reading span test scores, are more overt measures of reading proficiency. A third group effectively measures the respondent's attitude towards the given features or towards survey data in general.

If many of these factors impinge on our ability to read and interpret, it stands to reason that there will be a link between a proxy for the external "textual world," such as a corpus, and the sorts of answers respondents give on surveys. In the next section, we will consider how this relates to our own research data.

Corpus Data

For a number of years now, we have been looking at places in the Czech conjugational and declensional systems where a syntactic "slot" has multiple exponents whose usage is not clearly differentiated, a situation described variously as *competition* (Lečić, 2015), *variation* (Bermel & Knittl, 2012a, 2012b; Bermel et al., 2015a, 2017) or *overabundance* (Thornton, 2012).[4]

In common with other Slavic languages, Czech is highly inflected, and thanks to a series of far-reaching phonological changes over the last millennium, the conditions for deploying its broad assortment of inflectional material are not always clear (see Bermel & Knittl, 2012b, pp. 93–95 for a fuller discussion).[5] Consequently, while we are able to describe clearly for some syntactic slots what exponent is used there, for others there is considerable variation. Exponents may be described using a list-type approach ("the following lexemes use exponent A; others use exponent B") or using a collection of rules of thumb ("borrowings, multisyllabic stems, and labial consonant stems prefer exponent C; others prefer exponent D").[6] In addition to places where choice is clear-cut, there exists a transitional band of items where both exponents are used in some measure.

[4] An example of clearly differentiated usage is, e.g, between the exponents {em} and {ou} in the instr. sg.: the former is used with masc. and neut. nouns, while the latter appears with fem. nouns. The only place we get overlap—e.g, *s (v)okurkem ~ s (v)okurkou* 'with cucumber'—is where the gender of the noun is unstable across dialects. When usage is not clearly differentiated, often some factors or tendencies can be identified that contribute to choice, but none that clearly demarcate it.

[5] A further contributory factor to the persistence of variation in Czech may be the relatively weak position of the standard, which does not function as a common speech variety across the vast majority of the country (see, e.g, Sgall, 2011, p. 183, one among many texts that could be cited in this regard). Attempts at standardizing one or another variant tend to be perceived as applying only to formal written texts.

[6] Compare, for example, the appearance of fleeting [e] in the fem. and neut. gen. pl. and the description of the masc. animate nom. pl. exponents {i}~{ové}~{é} in Grepl et al. (1995), pp. 248–249, 256–257. The first is described in terms of a default form and the conditions under which insertion takes place, while the latter variation is described using overlapping semantic, phonological, and suprasegmental criteria that may apply. The same approach is used in the normative Internet Language Manual (Ústav pro jazyk český 2004).

In English, with its relatively impoverished inflectional morphology, the best higher-frequency environment in which to study this is the overlap between the so-called strong and weak verb classes in the past tense and the perfect, and it has been studied from various angles over the past several decades (Albright & Hayes, 2003; Bybee & Slobin, 1982; Chandler, 2010; Eddington, 2000; Haber, 1976; Prasada & Pinker, 1993, etc.).[7] In Czech, this overabundance is widespread across both verbal and nominal morphology (e.g, Bermel 2004a, 2004b, 2010; Bermel, Knittl, & Russell, 2015b); in particular, nominal morphology, with seven cases, two numbers, and between 10 and 15 major declension patterns for nouns, is a fertile area for the study of competition between variant forms.

Our research involved testing three such slots in Czech where this phenomenon occurs. Two of these are from the so-called hard masculine inanimate declension pattern (exemplar word *hrad* 'castle'). As a result of the merger and reorganization of the dominant o-stem class and the smaller u-stem class that had evidently already begun in proto-Slavic, in Czech the u-stem endings have spread widely across the old o-stem lexical stock in the genitive singular (gen. sg.) and the locative singular (loc. sg.), while the old o-stem endings have also penetrated the much smaller group of nouns that previously formed the u-stem class. The third is the result of a younger innovation in which feminine nouns inherited from the Proto-Slavic i-stem pattern (exemplar word *kost* 'bone') have acquired to a greater or lesser degree the exponents of the old Proto-Slavic ja-stem pattern (exemplar word *růže* 'rose') in the gen. sg. and most plural cases, forming a new pattern (exemplar word *píseň* 'song') whose membership is not all that clearly defined.

The Czech National Corpus

Our main interest was to see whether exposure had an impact on the way Czechs perceived these variant forms as well as how they used them. Our proxy for *exposure* was the Czech National Corpus (CNC), specifically the frequency with which forms occur in it.

By CNC, we mean specifically its layer of synchronic representative corpora of written language (SYN2000, SYN2005, SYN2010, and SYN2015).[8] Each of these corpora contain roughly 100 million tokens (excluding punctuation) and are *representative* in that they contain a mixture of text types, broken down at top level into *publicistika* 'journalistic texts,' *odborná* or *oborová literatura* 'specialist or non-fiction texts,' and *beletrie* 'imaginative texts.'[9] Attempts at producing *balanced* cor-

[7] Latinate nouns (*octopi~octopuses*, etc.) are another area where variation can be looked at in English, but it has been an area of more research in derivational morphology, where variation is more widespread (*normality~normalcy*, etc.). However, derivational morphology is not seen as having the same impact on our understanding of utterance structure and the creation of "grammatical" meaning as does inflectional morphology.

[8] On our proxies for *perception* and *use,* see the "Methodology" section below.

[9] This term is more often translated as "fiction," but in the CNC corpora prior to SYN2015, it

Table 2.1 Text-type breakdown (top level) in the SYN corpora

	SYN2000 (%)	SYN2005 (%)	SYN2010 (%)	SYN2015 (%)
Journalistic texts	60	33	33	33.33
Specialist texts	25	27	27	33.33
Imaginative texts	15	40	40	33.33

pora based on research into reading habits gave a variety of results, summarized in Table 2.1.[10]

It is hard to tell without access to the comparative research underlying these changes, but there is a clear shift in favor of a more equal balance of text types, simplifying the task of comparing results from various text types within the corpus.[11]

Our results drew on both the SYN2010 and SYN2005 corpora (Čermák et al. 2005; Křen et al. 2010). Our goal was to identify nouns that exhibit variation in usage in the cases targeted. We conducted targeted searches in SYN2005 using the corpus search engine to retrieve all word forms with a particular shape and grammatical tag, e.g, ending in <u> and tagged as a masc. inanimate gen. sg. noun, or ending in <a> with the same tag.[12] We then compared the resulting lists to find variant forms of a word, e.g, *jazyku/jazyka*, which represented the variation sought.

For each case, the lists of lemmas (with each ending and with both endings) ran to many thousands of items, so a manageable process was needed for verifying the data and catching potential errors. Our method is described in detail in Bermel and Knittl (2012b, pp. 97–98), but in brief: all concordances with the less frequent ending were verified manually, token by token, as were examples of the more frequent ending when it appeared in variation. We also removed all "nonwords" from the lists and looked at any errors in the lemmas, which are often a sign that mistagging may have occurred.

These measures did not remove all erroneous forms retrieved, which would have been a much larger job, but they eliminated a large number of them. Even so, the effect on our overall statistics was not all that evident: for most lexemes, the proportions remained roughly constant. We thus arrived at three lists of lexemes where there was variation between two forms in the cases in question.

includes examples of the genre *literatura faktu:* creative nonfiction such as memoirs, travelogues, etc.

[10] The latest corpus in the series, SYN2015, is not balanced in this fashion; see *inter alia* Čermák, Králík, and Kučera (1997) on the research underlying the original corpora and Cvrček, Čermáková, and Křen (2016) on the composition of SYN2015.

[11] A programmatic explanation for this shift away from "real-world balance" towards "text-type balance" is given in Cvrček et al. (2016).

[12] When lemmatization succeeds, the CNC always disambiguates and resolves in favor of one assignment for each place in the tag (unlike, for example, the Russian National Corpus, where ambiguities are never resolved and all possible tags are associated with a token). This disambiguation is partially rule-based and partially the result of a heuristic correction based on manual tagging of a portion of the corpus. When lemmatization fails, typically due to a very rare or poorly formed (misspelled) word form, no morphological analysis can take place and the form is tagged as *nerozpoznaný* 'unrecognized'; our searches will not have picked up such forms.

One early outcome of this work is that *variation* is a gradient feature. Looked at in absolute terms, we find variation with very high-frequency lexemes as well as very low-frequency lexemes. The proportion of case exponents in one vs. another form is also distributed along a scale: for one word, ending {1} may predominate, whereas for another word it might be ending {2}, and that dominance might be overwhelming or less strong. The only consistent observation is that few lexemes, other than those of low frequency or those where there is some sort of semantic motivation, exhibit equipollent distribution, e.g, both endings {1} and {2} occur in roughly even proportions. Where the variation is unmotivated or only partly motivated, there is almost always some sort of skew to the dominance of one exponent.

Over the past few years, we have used these lists, and a few others compiled in the meantime, to test various hypotheses about frequency. In particular, Bermel et al. (2017) demonstrated that proportional frequency of forms had a consistent effect, at least on the sort of tasks we were asking respondents to perform.

Using Corpus Data in Surveys

The nature of a survey using native-speaker respondents imposes limits on the amount of corpus data that we can test. Respondents fatigue easily; with a high number of short, repetitive tasks, we decided that we could not ask them to spend more than 15–20 min on the survey without risking their attention flagging. We had the advantage of being able to pay respondents, which proved a useful motivational tool, but even so, the number of factors we could include was constrained. In this round, then, we looked at proportional frequency only. It was operationalized by choosing lexemes that fell into one of six proportional bands. The first questions to address are: why use bands at all; why, if so, do we use six bands; and why were those particular boundaries selected for them?

What we are calling *bands* are often termed *bins*: all data found in a particular range is treated as having the same value. We might assume that the best option would always be to retain all precise values and thus not use any bands or bins: surely, it must be more precise to retain the information that lexeme C has exponent {1} 13.7% of the time, while lexeme D has exponent {1} only 12.5% of the time. However, retaining this level of precision has an impact on the way we test our data. It implies a level of precision that in the real world may not exist, i.e, that because a 100-million-word corpus has those particular values, a native speaker will be more likely to favor exponent {1} in lexeme C than exponent {1} in lexeme D, and will be correspondingly more likely to use it in the first scenario than the second. For this reason, tests using bins may prove to be more realistic if we believe that corpora are best interpreted as a rough guide to the linguistic environment rather than an exact one; and that our abilities to track this linguistic environment may be approximate rather than precise.

To reduce at least one aspect of uncertainty, we limited our choice of nouns to those where at least 100 tokens in the case in question were found in a 100-million-token corpus (1 ipm). While this is admittedly an arbitrary level, we felt that it was necessary to ensure the validity of results. A set with four tokens of exponent {1} and two tokens of exponent {2} gives a proportional frequency of 67%:33%, but if only two tokens had been different, the proportions would have been reversed. With a sample of $N \geq 100$, the chance of this happening is correspondingly reduced.

We set the number of bands and the particular boundaries between them opportunistically. For us, the most important criteria were that we get enough granularity in the results to be able to draw clear conclusions, and that we draw the boundaries around our bins in such a way that each of them represents a meaningful number of items. If we create a bin with few or no items in it, the information it yields will be limited and we will have a severely constrained choice of lexemes to use in our survey. In other words, we are not proposing that these specific bands have any inherent meaning themselves, i.e, that using six bands instead of seven indicates a rougher granularity of response overall, or because a word falls into the fifth instead of the sixth band that its behavior is qualitatively different. Instead, we are testing the usefulness of a scale itself: whether the proportional frequency of items in the linguistic environment makes a difference to people's judgments and choices.

For our purposes, then, the most important feature of a scale is that the bands each contain adequate numbers of lexemes for us to construct a survey, and that the survey contain enough levels to assess the variation properly. How we assess the variation has an effect on (and is affected by) the statistical measures chosen.

Previously, for example, we had experimented with seven bands and four bands. The latter had little granularity and thus results were not as clear as we had hoped, while the former presupposed a "central" band with roughly equal proportions of each exponent—which, as it turned out, were very difficult to find. This is because, as mentioned in the section "The Czech National Corpus," unmotivated and partially motivated variation tends to result in a skew dominance, where one exponent predominates in the vast majority of circumstances. In other words, where a firm criterion for choosing one form over another is lacking, frequency itself becomes a criterion, with users perceiving one form as "default" or "normal" and the other as "rare" or "unusual" to varying degrees. In the end, we went with a division into six unequally sized bands that allowed us a reasonable choice of lexical items for each band. The middle two bands were much broader (35% each), while the outside bands were very narrow (1% each), as this is where we find the greatest number of lexemes with variant forms.

We further restricted our choice of lexemes by checking our findings in both SYN2005 and SYN2010, two corpora with identical high-level structures (see Table 2.1 above). To warrant inclusion in our survey, a lexeme had to fall into the same proportional frequency band in both corpora. The resulting set of nouns can be seen in Table 2.2.