Illuminating variation

Individual differences in entrenchment of multi-word units

This dissertation presents research into variation between and within participants in their metalinguistic judgments about, and processing of, multi-word sequences. It thus contributes to the development of the usage-based framework in linguistics. Individual differences in mental representations of language naturally follow from a usage-based approach. Since people differ in their linguistic experiences, they are expected to differ in the extent to which a linguistic construction is entrenched in their mental lexicons. Furthermore, a language user gains new linguistic experiences over time, and mental representations of language are hypothesized to change accordingly. There is a shortage of empirical data on these types of variation, though.

To examine inter- and intra-individual variation, two studies in this dissertation use a test-retest design: participants performed the same judgment task twice within the space of a few weeks. In another study, recruiters, job-seekers, and people not (yet) looking for a job performed a completion task, a voice onset time task, and a metalinguistic judgment task consecutively. These groups differ in their exposure to a particular register (job ads), which is expected to lead to differences in mental representations of language.

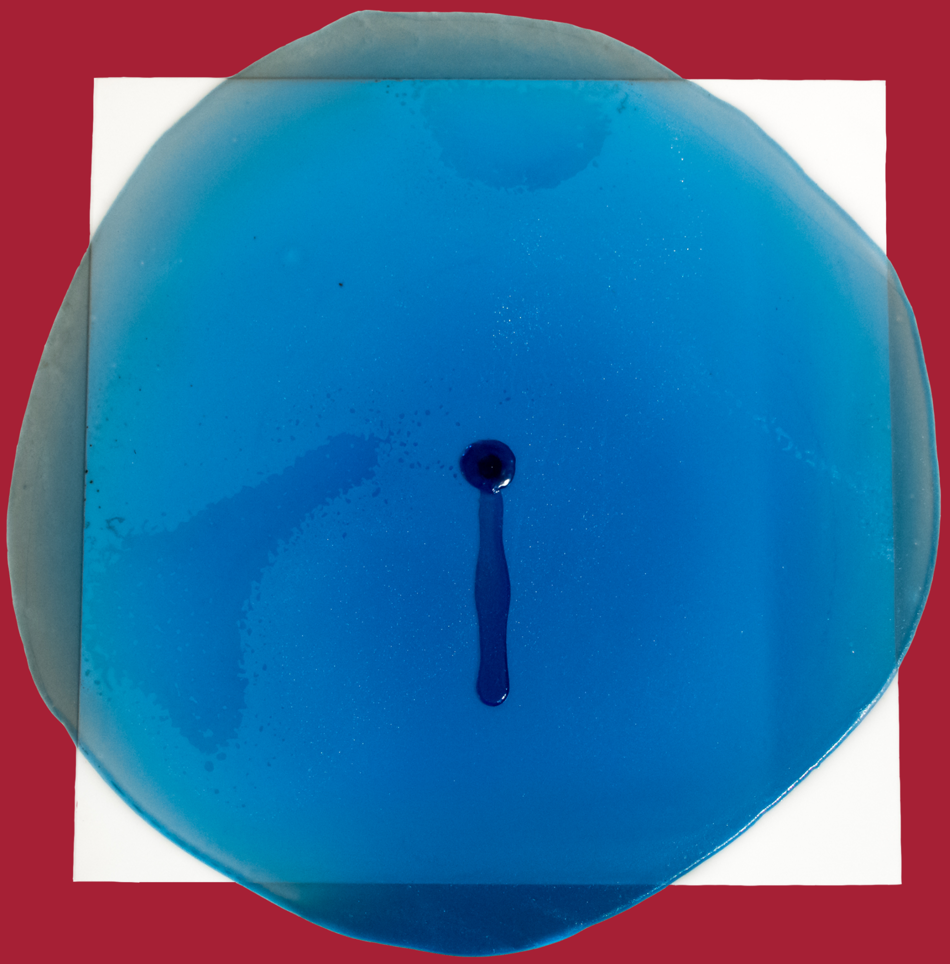
Véronique Verhagen compares participant-based measures and measures based on amalgamated data of different people (corpus-based frequencies, surprisal, cloze probabilities) as predictors of performance in psycholinguistic tasks. This provides insight into individual variation and the merits of going beyond amalgamated data. The thesis demonstrates how investigations of inter- and intra-individual variation in psycholinguistic data advance our understanding of the dynamic character of mental representations of language.

ISBN 978-94-6093-333-2

:LOT Netherlands Graduate School of Linguistics
Landelijke Onderzoekschool Taalwetenschap

:LOT
548

Véronique Verhagen **Illuminating variation**



Véronique Verhagen

Illuminating variation

Individual differences in entrenchment of multi-word units

:LOT

Tilburg University

Illuminating variation

Individual differences in entrenchment of multi-word units

Published by
LOT
Kloveniersburgwal 48
1012 CX Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
http://www.lotschool.nl

Cover illustration: picture of an artwork by Piet Stockmans, photographed by Daphne Snijders. To me, it visualizes the dynamic character of mental representations of language, which may best be viewed as moving targets.

ISBN: 978-94-6093-333-2
NUR: 616

Illuminating variation

Individual differences in entrenchment of multi-word units

Proefschrift

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus, prof. dr. K. Sijtsma,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op vrijdag 10 januari 2020
om 13.30 uur

door

Véronique Anne Yvonne Verhagen

geboren op 12 december 1985 te Eindhoven

Promotor	prof. dr. A.M. Backus
Copromotores	dr. M.B.J. Mos
	dr. J. Schilperoord
Promotiecommissie	prof. dr. W.B.T. Blom
	prof. dr. E. Dąbrowska
	prof. dr. H.-J. Schmid
	dr. E. Zenner

The research reported in this dissertation was supported by an NWO grant (*de Nederlandse organisatie voor Wetenschappelijk Onderzoek* – the Dutch Research Council, project number 322-89-004)

Voorwoord

Het allereerste college dat ik volgde als student, was dat van het vak Taalwetenschap. Het allereerste college dat ik als docent gaf, was het college Taalwetenschap, en het vond plaats in de zaal waar ik destijds mijn eerste college had gevolgd (het handboek en de opdrachten waren overigens niet meer hetzelfde – ik wil niet de indruk wekken dat er geen ontwikkeling plaatsvindt in deze faculteit, in tegendeel!). In de jaren die daarop volgden, heb ik ook aan de Universiteit Leiden en bij de lerarenopleiding Nederlands aan Fontys taalkundevakken gedoceerd. Die activiteiten hebben het afronden van mijn promotieonderzoek 'ietwat' vertraagd, maar ze hebben ook me veel waardevolle kennis, ervaringen, en contacten opgeleverd. Ik ben dankbaar voor de mogelijkheden die mij in dat opzicht zijn geboden. Minstens zo dankbaar ben ik voor de ondersteuning van mijn begeleiders bij het voltooien van mijn proefschrift.

Om te beginnen Maria; zonder haar voortvarendheid en betrouwbaarheid was deze dissertatie er wellicht wel gekomen, maar dan had het gegarandeerd langer geduurd. Dankjewel voor je betrokkenheid en goede adviezen, en je fijne gezelschap tijdens conferenties. Na een workshop in Potsdam vroeg Jon Sprouse of wij misschien zussen waren. Jij antwoordde toen verbaasd *Nee,* en voegde er aan toe: *hoogstens 'academic sisters'.* Je bent de beste grote academische zus die ik me kan wensen.

Ad ben ik zeer dankbaar voor zijn nimmer aflatende vertrouwen. Er zijn niet veel hoogleraren die zo wijs, ruimhartig, en *in touch* met hun *feminine side* zijn als jij. De hoeveelheid mensen die een beroep op je doen is onvoorstelbaar groot en toch neem je altijd de tijd voor alle vragen die iemand heeft. Als ik promovendi ontmoette die Ad kenden, waren ze steevast jaloers op het feit dat hij mijn promotor was.

Joost bewonder ik om zijn mooie invallen en formuleringen, en dank ik voor zijn aanmoedigingen om te "ronken en blazen" en zijn vermogen om zaken vanuit een andere hoek te bezien. Tijdens de verdediging van mijn masterscriptie vroeg je mij: *En als je het omgedraaid had? Als je mensen had gevraagd te beoordelen hoe weínig de woorden bij elkaar horen?* – een mogelijkheid die nooit in mij was opgekomen. Ook tijdens mijn promotieonderzoek kwam je telkens met waardevolle voorstellen om zaken eens om te draaien en wees je mij op het moois in mijn data als ik vooral gefocust was op wat we er níet mee konden aantonen.

Antal van den Bosch ben ik zeer erkentelijk voor zijn waardevolle adviezen en het feit dat hij mij in contact heeft gebracht met Jakub Zavrel en Louis Onrust. Jakub is de oprichter van Textkernel – een bedrijf dat gespecialiseerd is kunstmatige intelligentie op het gebied van HR en recruitment. Eén van hun instrumenten, Jobfeed, zoekt het internet af naar vacatures. Dankzij deze technologie en de behulpzaamheid van Jakub en zijn collega's, heb ik een corpus met vacatureteksten tot mijn beschikking gekregen. Mijn dank is groot. Louis' hulp bij het analyseren van de dataset bestaande uit ruim 1,36 miljoen vacatureteksten was van onschatbare waarde. Ik ben hem heel dankbaar voor zijn geduld en generositeit.

Prof. dr. Blom, prof. dr. Dąbrowska, prof. dr. Schmid, and dr. Zenner, thank you very much for accepting the invitation to be part of the committee. I am greatly honored that you have read my work and that you are willing to discuss it with me.

Als promovenda en beginnend docent heb ik deel mogen uitmaken van een departement dat gekenmerkt wordt door een buitengewone mate van kwaliteit en collegialiteit. Adriana, Alex, Alwin, Anne, Annemarie, Carel, Charlotte, Chris, Christine, Constantijn, David, Diana, Debby, Emmelyn, Emiel, Emiel, Eriko, Fons, Hans, Jacqueline, Jan, Jan, Janneke, Jorrig, Jos, Joost, Julie, Juliette, Karin, Kiek, Lauraine, Leonoor, Lieke, Loes, Mandy, Marc, Maria, Marie, Mariek, Marieke, Marije, Marjolein, Marlies, Martijn, Martin, Menno, Monique, Nadine, Nadine, Naomi, Neil, Nynke, Paul, Per, Peter, Rein, Renske, Ruben, Ruud, Saar, Sander, Tess, Yan, en Yevgen, dank jullie wel voor alle interessante gesprekken, de fijne samenwerking in onderwijsactiviteiten, het medeleven toen redacteur R. mij tot wanhoop dreef, de verkwikkende wandelingen in de Oude Warande, de geweldige optredens van de Malle-band, de fantastische departementsuitjes, Sinterklaasgedichtjes, en kerstdiners.

Voordat ik als promovenda aan de slag ging, ben ik als student gevormd door het werk van Ad, Carine, Erna, Karen, Guus, Helma, Jan, Jan Jaap, Jeanne, Jos, Kutlay, Leon, Max, Mia, Odile, Piia, Rian, Sander, Sjaak, Ton, en Tineke. Dank voor de boeiende colleges die ik met veel interesse bij jullie heb gevolgd en voor het feit dat ik 'op kamers' mocht op de 4^{e} verdieping.

Naast mijn aanstelling als onderzoeker in Tilburg, heb ik gedurende anderhalf jaar taalkundevakken mogen verzorgen in Leiden bij de opleidingen Nederlandse taal en cultuur en Taalwetenschap. Alex, Arie, Esther, Gijsbert, Maaike Beliën en Maaike van Naerssen, Maarten, Olga, Ronny, Roosmaryn, Saskia, Tanja, Ton, en Vivien, dank jullie wel voor deze leuke en leerzame tijd.

Terwijl ik mijn proefschrift aan het afronden was, ben ik bij Fontys gaan werken bij de lerarenopleiding Nederlands. Arina, Bart, Bas, Chantall, Claudia, Elly, Esther, Gerbert, Hanneke, Henriëtte, Jan, Julia, Kristien, Maartje, Maartje, Margriet, Monica, Nanette, Petra, en Rudie, dank voor het mij verwelkomen en wegwijs maken in een wereld die nieuw is voor mij. Dank ook voor jullie interesse ten tijde van het inleveren van het manuscript en het delen in de vreugde toen ik bericht van de commissie ontving.

Tot slot wil ik mijn lieve en leuke familie en vrienden bedanken voor het deelnemen aan experimenten, het vragen én het niet vragen naar de voortgang, het meedenken over de lay-out en de kaft, en nog meer voor de vele mooie, grappige, bijzondere niet-proefschriftgerelateerde momenten. Een speciaal woord van dank aan mijn ouders, wier betrokkenheid en zorgzaamheid oneindig groot is.

Contents

Chapter 1 Introduction

Suppose a number of people encounter the utterance *Bij gelijke geschiktheid gaat onze voorkeur uit naar een vrouwelijke kandidaat* ('In case of equal qualifications, we will give preference to female candidates'), to what extent would they differ in the linguistic units they employ in processing it, and can we explain these differences? For a long time, linguists have regarded words and grammatical rules as the basic units in language. However, it has become increasingly clear that this is not sufficient as a description of how language is organized in our minds, as there is considerable evidence that we have a much more varied set of linguistic units at our disposal. While an utterance such as *Bij gelijke geschiktheid gaat onze voorkeur uit naar een vrouwelijke kandidaat* could be produced and understood by accessing the individual words and the syntactic structure in which they are embedded, speakers may also employ larger processing units. They can, for example, make use of multi-word units (e.g. *bij gelijke geschiktheid*) and partially schematic units (e.g. *gaat* ART/POSS *voorkeur uit naar* NP). As psycholinguistic research has uncovered, some of these chunks of language are processed more quickly, recalled more easily, and deemed more familiar than others. This suggests that they differ from each other in representational strength, or, put differently, in degree of entrenchment. Usage frequency appears to play a key role in the process of entrenchment: the more a linguistic unit is used, the more it becomes entrenched in the speaker's mental lexicon, thus making it easier for this speaker to retrieve and process it.

If usage-based models of linguistic representations are correct in positing such a strong link between usage frequency and entrenchment, it follows that the extent to which a linguistic unit is entrenched varies from person to person, as well as over time. There is a shortage of empirical data on these types of variation, though. As I will discuss in more detail in Section 1.1.1 and in the following chapters, the past five decades have seen a wealth of studies yielding evidence in support of usage-based theories of language acquisition and processing, but these studies have paid little attention to inter- and intra-individual variation. A central aim of the studies presented in this dissertation is to demonstrate that insight into these types of variation is a prerequisite for a veridical description of mental representations of language. The studies thus aim to contribute to usage-based theories of language by examining variation in entrenchment of multi-word units.

1.1 Usage-based linguistics

Linguistic theories ought to posit a model of linguistic knowledge that explains that speakers can produce and understand an infinite number of utterances, that also accounts for the ease and speed with which speakers are able to process language, and that is learnable. Usage-based linguistics is a framework that accounts for productivity, real-time processing, and learnability by envisioning linguistic knowledge as dynamic networks of constructions which are shaped by the cognitive response to social behavior, thus accommodating insights from both psycholinguistics and sociolinguistics. In this framework, mental representations of language consist of form-meaning pairings (i.e. constructions) that are taken to emerge from, and are continuously shaped by, experience with language together with general cognitive skills and processes such as categorization, schematization, and chunking (Barlow & Kemmer 2000; Bybee 2006; Goldberg 2006; Tomasello 2003; A. Verhagen 2005). Linguistic constructions vary in size –ranging from single morphemes (e.g. *like*) to multi-word units (e.g. *to all intents and purposes*)– and in schematicity –ranging from lexically specific constructions (e.g. *equal qualifications*) to partially schematic (e.g. V-*able*) and fully schematic ones (e.g. SUBJECT VERB DIRECTOBJECT). The fact that, on a usage-based account, language use continuously shapes mental representations of language makes that linguistic constructions are entrenched to varying degrees.

1.1.1 Degrees of entrenchment

Entrenchment can be defined as "the degree to which the formation and activation of a cognitive unit is routinized and automated" (Schmid 2007:119; see also Langacker 1987). Frequency of use is taken to be a key factor determining degree of entrenchment. The more frequently a speaker encounters and uses a particular linguistic structure, the more the mental representation of this structure will become entrenched. As a result, it can be activated and processed more quickly, which, in turn, increases the probability that this form is used to express the given message, making this construction even more entrenched. Conversely, extended periods of disuse weaken the representation (Langacker 1987: 59).

An impressive body of research shows that people are very much attuned to frequency in language. We are sensitive to distributional properties of sound sequences, morphemes, words, word sequences, and syntactic patterns, and we make use of this information in language acquisition and processing (for overviews see Diessel 2007; N. Ellis 2002; Gries & Divjak 2012; Saffran 2003). With regard to multi-word units –the type of construction that I focus on in my studies– numerous studies have demonstrated a strong relationship between the frequency with which a word sequence occurs in the language and the extent to which its formation and activation in the minds of speakers is routinized, as evidenced by pronunciation duration and phonological reduction (e.g. Arnon & Cohen Priva 2013; Bannard & Matthews 2008; Bybee & Scheibman 1999; Janssen & Barber 2012), perceptual identification (e.g. Caldwell-Harris, Berant & Edelman 2012), reading times (e.g. N. Ellis & Simpson-Vlach 2009; Fernandez Monsalve et al. 2012; McDonald & Shillcock 2003; Siyanova-Chanturia, Conklin & van Heuven 2011; Smith & Levy 2013), phrasal decision times (e.g. Arnon & Snider 2010; Jolsvai, McCauley & Christiansen 2013), and N400 effects (e.g. Frank et al. 2015). These findings suggest that linguistic constructions vary in the extent to which they are entrenched in speakers' mental constructicons and that degree of entrenchment is strongly correlated with usage frequency.

As Tomasello (2007: 282, as cited in Divjak 2016) aptly remarks, "[t]oday, very few linguists would seriously deny the existence of frequency effects in language. The real argument within linguistics is how far these effects go". I propose that an investigation of inter- and intra-individual variation in psycholinguistic data can advance our understanding of the effects of usage frequency on language processing and mental representations of language. These kinds of variation naturally follow from a usage-based perspective. In order to do justice to the usage-based approach, researchers ought to attend to such variation, examine to what extent it is usage-based and what it reveals about the dynamic nature of mental representations.

1.1.2 Variation in degrees of entrenchment

If representational strength is determined largely by usage frequency, there are likely to be differences in entrenchment across individuals, even within a group that is relatively homogeneous in terms of sociolinguistic characteristics, since language users differ in their linguistic experiences. It is not known, though, how large these differences are. Given that speakers are able to communicate rather successfully, it appears that linguistic representations do not diverge widely. Still, differences may be more profound than is often assumed. While sharing knowledge of high-frequency schematic structures (e.g. the transitive construction SUBJECT VERB DIRECTOBJECT) and a large inventory of specific linguistic elements such as single words and multi-word chunks, speakers differ in the extent to which they encounter and use particular words, word combinations, and (partially) schematic constructions. The frequency with which they experience such constructions differs, the contexts in which they encounter them differ, and the ways in which they combine various constructions differ as well. Such differences are expected to result in variation across speakers in linguistic representations.

In addition to inter-individual variation, a usage-based approach predicts intra-individual variation. Effects of usage on linguistic knowledge are not restricted to children acquiring their mother tongue(s) and adults acquiring a foreign language; they also hold for adult native speakers. All language users gain new linguistic experiences throughout their lives, and usage-based linguistics predicts mental representations of language to change accordingly.

To date, few studies have examined the variability of mental representations of language in adult native speakers. Cognitive linguists often make use of corpus data; these corpora are usually an amalgamation of texts and/or recordings of spoken language from many different language users, which are unlikely to be fully representative of the linguistic experiences of the people taking part in a study and unlikely to be equally representative for all participants alike. Some researchers have analyzed corpora composed of data of an individual speaker (e.g. Barlow 2013; Dąbrowska 2014; Schmid & Mantlik 2015). Their findings point to individual differences in the use of various constructions. However, patterns of use as observed in corpus data cannot be equated with the degrees to which constructions are entrenched in the mind of the speaker. In order to link these patterns of use in corpus data to entrenchment, they need to be supplemented with data from psycholinguistic experiments.

While it is starting to become common practice to analyze experimental data by means of statistical models that account for individual differences (e.g. mixed-effects models), the variation present in psycholinguistic data is rarely analyzed in its own right. Experimental data are usually reported as aggregated scores, without regard for the degrees of variation and the information they may convey. Furthermore, whenever a study involves multiple types of experimental tasks, these are commonly conducted with different groups of participants. Consequently, variation across tasks and variation across speakers are confounded. As a result, such studies yield little insight into inter-individual variation. In addition, participants are seldomly asked to perform a task multiple times. Therefore, not much is known about the degrees of intra-individual variation from one moment to another.

1.2 Multi-word units

In this dissertation, I focus on multi-word units as linguistic constructions. In the last couple of decades, the importance of multi-word units in language acquisition and processing has come to the fore. Analyses of the utterances produced by 2- and 3-year-olds and the input they had received reveal that children stick close to word strings they have encountered in the input (Dąbrowska & Lieven 2005). In addition, experimental research has shown that the more frequently phrases occur in child-directed speech, the better children are at processing and (re)producing them (Arnon & Clark 2011; Bannard & Matthews 2008; McCauley & Christiansen 2014). These lexically specific constructions form the basis for schematic constructions; by generalizing over specific instances, children are able to arrive at more abstract schemas (Goldberg 2006). The emergence of schematic constructions does not imply that multi-word units become less important. In fact, usage-based theories consider more specific constructions as more basic:

> lower-level schemas, expressing regularities of only limited scope, may on balance be more essential to language structure than high-level schemas representing the broadest generalizations. (…) For many constructions, the essential distributional information is supplied by lower-level schemas and specific instantiations (Langacker 2000: 30-31).

Syntactic and semantic analyses of instances of various constructions provide support for this point of view (e.g. A. Verhagen 2003). This is complemented by empirical evidence that indicates that adult speakers store phrases and that the use of these ready-made chunks facilitates sentence comprehension and production (e.g. Arnon & Snider 2010; Arnon & Cohen Priva 2013; Bybee & Scheibman 1999; Caldwell-Harris, Berant & Edelman 2012; Dąbrowska 2014; N. Ellis & Simpson-Vlach 2009; Janssen & Barber 2012; Jolsvai, McCauley & Christiansen 2013; Shaoul, Baayen & Westbury 2014; SiyanovaChanturia, Conklin & van Heuven 2011; Tremblay & Baayen 2010). This has led cognitive linguists to the viewpoint that the use of ready-made chunks is the basic mode of using language (e.g. Bybee 2007: 279-280; Dąbrowska 2014: 642; Wray 2002, also see Christiansen & Chater 2008, 2016 and McCauley, Isbilen & Christiansen 2017).

1.3 This dissertation

The studies presented in this dissertation examine variation between and within participants in their metalinguistic judgments about, and processing of, multi-word sequences. They investigate the variation present in the data and the extent to which this variation can be considered meaningful. From a theoretical perspective, insights into the degree of individual variation contribute to a refinement of usage-based accounts. Findings indicate to what extent variation should be part of linguistic descriptions. They also enable us to delineate more precisely the limitations of different research methods that aim to tap into degrees of entrenchment.

This dissertation also serves as a proof of concept. The studies employ research designs and methods that are well suited to test hypotheses that follow from usage-based theories of linguistic knowledge and language processing, and to yield insight into inter- and intra-individual variation. The approach adopted here can be extended, in future research, to other groups of speakers, other linguistic registers, and other types of linguistic constructions. In this dissertation, multi-word units are the construction of interest, since they have been shown to play a pivotal role in language processing. Another reason to focus on multi-word sequences is that this type of construction lends itself well to the investigation of usage-based variation. Registers and social groups are likely to differ more notably in the usage of multi-word units than in experience with schematic constructions. Schematic constructions have a more general and abstract meaning than lexically specific constructions. As such, schematic constructions may be less sensitive to differences in usage contexts that differ from one person to another. In Chapter 7, I discuss to what extent the findings presented in this dissertation can be expected to hold for constructions other than multi-word units.

1.3.1 Outline

Chapters 2 through 5 report on experimental research combining corpus analyses and psycholinguistic data. In Chapter 6, I reflect on the methodological lessons that can be learned from these studies; in Chapter 7, I discuss the theoretical implications. Chapters 2, 3, 4, and 6 are based on articles published or submitted for publication in peer-reviewed journals.

Chapters 2 and 3 present two studies that examine inter- and intra-individual variation in metalinguistic judgments. The latter is investigated by means of a test-retest design: participants performed the same task twice within the space of one to three weeks. In both studies, participants were asked to assign familiarity ratings, using the method of Magnitude Estimation, to a set of prepositional phrases that cover a wide range of corpus frequencies. In Chapter 2, these phrases were presented in isolation as well as in a sentential context, to investigate whether context affects perceived degree of familiarity and inter- and intra-individual variation in judgments. The judgment task in Chapter 3 involved isolated phrases only. In this study, participants used either a 7-point Likert scale or a Magnitude Estimation scale. The research design employed in Chapter 3 thus yielded data on variation across items, across participants, across time, and across rating methods.

Chapters 4 and 5 report on three experiments that were conducted with three groups of participants: recruiters, job-seekers, and people not (yet) looking for a job. These groups can be expected to differ in experience with word sequences that typically occur in job ads (e.g. *goede contactuele eigenschappen* 'good communication skills'); they are not expected to differ systematically in experience with word sequences characteristic of news reports (e.g. *de Tweede Kamer* 'the House of Representatives'). The participants first performed a completion task, which offers insight into the expectations people generate about upcoming words. This was followed by a voice onset time (VOT) task, which provides data on the speed with which the participants process the word strings. After that, the participants assigned familiarity ratings to the word sequences using Magnitude Estimation. Chapter 4 reports on the completion task and the VOT task; Chapter 5 reports on the metalinguistic judgment task.

In Chapters 4 and 5, I examine the relationship between amount of experience with a particular register and (i) the expectations people generate about upcoming words when faced with word strings characteristic of that register; (ii) the speed with which they process such word strings; and (iii) how familiar they consider these word strings to be. Furthermore, I investigate the relationships between data elicited from an individual participant in different types of psycholinguistic tasks using the same stimuli. Comparisons of participant-based measures and measures based on amalgamated data of different people as predictors of performance in psycholinguistic tasks provide insight into individual variation and the merits of going beyond amalgamated data.

Chapter 6 highlights the merits of multi-method research in linguistics and offers an overview of key considerations in the design of such research. Chapter 7, finally, provides a summary of the main findings and discusses the theoretical implications as well as suggestions for future research.

Chapter 2

Abstract

Judgments are often used in linguistic research. Not much is known, however, about the variation of such judgments within and between participants. From a usage-based perspective, variation might be expected: with judgments based in representations, and representations resulting from input and use, both inter- and intra-individual variation are likely. This study investigates the reliability of metalinguistic judgments, more specifically familiarity judgments, for Dutch prepositional phrases (e.g. op de bank, 'on the couch'). Familiarity judgments for 44 PPs offered in isolation and in a sentential context were given by 86 participants in two identical test sessions, using Magnitude Estimation. Aggregated scores (averaged over participants) are remarkably consistent (Pearson's r = .97), and in part predicted by corpus frequencies. At the same time, there is considerable variation between and within participants. Context does not reduce this variation. We interpret both the stability and instability to be real reflections of language: a relatively stable system in a speech community consisting of speakers who are variable and forever changing. The results suggest that judgment data are informative at different levels of granularity. They call for more attention to individual variation and its underlying dynamics.

This chapter is based on:

Verhagen, V. & Mos, M. (2016). Stability of familiarity judgments: Individual variation and the invariant bigger picture. *Cognitive Linguistics, 27*(3), 307–344. https://doi.org/10.1515/cog-2015-0063

Acknowledgements

I thank Dominique Mellema for his help in collecting the data, and Dagmar Divjak and two anonymous reviewers for their helpful comments on the paper we submitted to Cognitive Linguistics.

Chapter 2 Stability of familiarity judgments: individual variation and the invariant bigger picture

2.1 Introduction

Metalinguistic judgments constitute an oft-used type of data in a variety of fields within linguistics, ranging from grammaticality and acceptability judgments (e.g. Sprouse & Almeida 2012 for syntactic patterns; N. Ellis & Simpson-Vlach 2009 for formulaic language; Granger 1998 and Gries & Wulff 2009 for collocations and constructions in L2 speakers) to judgments regarding productivity (e.g. Backus & Mos 2011) and idiomaticity (e.g. Wulff 2009). Various researchers have criticized the validity and reliability of metalinguistic judgments (e.g. Bornkessel-Schlesewsky & Schlesewsky 2007; Sampson 2007). Still, the general assumption behind the use of judgment data in linguistic research is that they provide us with information about linguistic representations, overlaid with certain amounts of processing difficulty, depending on the specifics of the task and the setting, that cannot be deduced from natural language use or psycholinguistic, experimental data. All the more remarkable is the fact that we do not know how stable and therefore reliable such judgments are. Already in 1987, Labov stated: “The most obvious hiatus in the foundations of modern linguistics is the absence of a concern for the reliability and validity of the introspective judgments that form the main data base of grammatical research”.

Since Labov’s observation, several decades have passed and still the reliability of metalinguistic judgments has not been investigated thoroughly. To be sure, there is a large body of literature on ratings (for an overview see Schütze & Sprouse 2013) and various studies have compared judgment data to other types of data such as expert intuitions (Dąbrowska 2010), textbook classifications (Sprouse & Almeida 2012), and corpus data (Balota et al. 2001). However, such comparisons do not provide conclusive evidence about the stability of and variation in judgments. Typically, judgments by different participants are averaged and inter-individual differences are regarded as ‘noise’ (but not always, viz. Dąbrowska 2012, Dąbrowska 2013; Barlow 2013; Barth & Kapatsinski 2014).

Given that people differ in their linguistic experiences and in the language they produce themselves, individual differences actually are to be expected in judgment data (depending on the items that are judged, a point to which we will return below). A discrepancy between one person’s judgments and those of other people, or between someone’s judgments and corpus data, does not necessarily invalidate these judgments. People may differ from each other in real and meaningful ways, each expressing their own linguistic representations. The most thorough and direct way of examining the stability of judgments, while allowing for differences between individuals as well as between items, is to have people judge the same linguistic stimuli several times, which is not common practice.

In this paper, we address the issue of variability in linguistic judgments. The paper starts by introducing the particular type of stimulus items and judgment used in the current study: familiarity ratings for multi-word units. We argue where and why differences between people (hereafter *inter-individual variation*) as well as within a single language user (*intra-individual variation*) might be expected. This is followed by a discussion of an important factor that could influence these two types of variation: providing a context to stimulus items. We then report on the outcomes of an experimental study into the stability of metalinguistic judgments and the relationship between these judgments and corpus data. We argue how the observed stability and instability in judgments could be accounted for in a usage-based framework and how it calls for further investigation of the variability of (meta)linguistic representations. As such, this study contributes to our understanding of the relation between individuals' judgments on the one hand and their linguistic representations as well as the entrenchment of patterns in the speech community on the other.

2.1.1 Judging multi-word units

In this study we focus on multi-word units, and the judgment data concern the perceived familiarity of these units. A multi-word unit is a string of words that are taken to be stored together, as a whole, in one's linguistic repertoire (a.o. Wray 2002). Multi-word units have characteristics that make them suitable to be assessed in a familiarity judgment task. They are small enough to be stored as chunks. Moreover, they are plausible units as they form a semantic and syntactic unity. This also means that it is easier for people to provide familiarity ratings for multi-word strings than for entire sentences, skip-grams (i.e. discontinuous multi-word *n*-grams, such as *go to … lengths*) or bound morphemes.

The basis for the entrenchment of multi-word sequences is the fact that words tend to occur in certain constructions and collocate to form multi-word units (Stefanowitsch & Gries 2003 and many others). Numerous studies provide evidence that language users are sensitive to the likelihood of words to co-occur (e.g. Jurafsky et al. 2001; Mos et al. 2012). If one takes a usage-based perspective on language processing and representation, as we do here, distributional patterns are inextricably related to one's cognitive linguistic representations, as knowledge of a language is in large part built from (mostly implicit) memories of past linguistic experiences (see, for example, J. Taylor 2012). To put it more precisely, our linguistic representations emerge from our experience with language —that is, the language we encounter and produce ourselves— together with general cognitive skills and processes such as schematization, categorization and chunking. The latter, of particular importance here, is the process “by which sequences of units that are used together, cohere to form more complex units” (Bybee 2010: 7). ‘Complex’ here means that the unit consists of multiple elements that are packaged together in cognition. The process of chunking is thought to occur in adults as readily as in children, and applies to all kinds of sequences of linguistic elements.

The principal experience that triggers chunking of multi-word sequences is frequency-based: repetition (Bybee 2010). The more a sequence of words is used together, the more entrenched it becomes as one chunk. An impressive body of research has revealed a log-linear relationship between usage frequency –usually estimated on the basis of corpus data— and processing as measured in psycholinguistic experiments (see for instance N. Ellis 2002; Diessel 2007). Furthermore, log-transformed frequency scores have been shown to resemble the way language users perceive differences in frequency (e.g. Popiel & McRae 1988 for idioms; Balota et al. 2001 for single words).

These studies, however interesting, do not tell us much about variation in individuals’ cognitive representations of multi-word sequences —that is, the synchronic result of accumulated exposure and chunking— nor about people’s ability to reliably report on these representations. In order to investigate the perceived degree of ‘chunkiness’ of a word sequence we designed a set of prepositional phrases and asked people to judge these phrases twice within the space of a few weeks (a more detailed description is given in Section 2.2 below). Participants were asked to provide familiarity judgments. Familiarity of a word sequence (or any other type of linguistic element) is taken to rest on frequency and similarity to other words, constructions or phrases (e.g. Bybee 2010: 214). As such, familiarity taps into exposure and chunking, while it does not require introducing a new concept to participants. Asking participants to provide ratings for ‘familiarity’ rather than ‘entrenchment’, ‘chunkiness’, or ‘unit status’ means that it is not necessary to introduce jargon. Furthermore, it does not evoke a right/wrong distinction, and the concept of familiarity involves both one’s own usage and one’s experience with other people’s use of the items.

A substantial number of studies have made use of familiarity ratings for words, word pairs, phrases, idioms, and metaphors. These ratings were found to be significant predictors of reading times (e.g. Cronk et al. 1993; Juhasz & Rayner 2003; Williams & Morris 2004), as well as performance on lexical decision and speeded naming tasks (e.g. Gernsbacher 1984; Connine et al. 1990; Blasko & Connine 1993; Juhasz et al. 2015), speeded semantic judgment tasks (among others, Tabossi et al. 2009), and perceptual identification tasks (Caldwell-Harris et al. 2012). Gernsbacher (1984: 227) states that asking participants to rate how familiar they are with a word is a simple tool for collecting a measure of the extent and type of previous experience respondents have had with each word. Juhasz et al. (2015: 1005), in like manner, write: "Rated familiarity can be thought of as a measure of subjective frequency such that it indexes the experience that an individual has with a given word." As familiarity crucially depends on prior linguistic experiences, it implies variation, both across speakers and over time. These two types of variation are discussed in more detail successively.

2.1.2 Inter- and intra- individual variation

People differ, from one person to the next, in the way in which, and the frequency with which, they encounter and use particular word strings. As J. Taylor (2012: 250) puts it: "It is evident even to the most casual observer that speakers of the 'same' language may exhibit variation in their usage patterns according to their geographical provenance, their social status, their educational background, their age, gender, ethnicity, and so on". If linguistic representations are assumed to be based on one's linguistic experiences, such differences are expected to give rise to variation in these linguistic representations.

Within the Cognitive Linguistics framework, the idea that people may differ considerably in their linguistic knowledge, not just at the level of lexical repertoires, has been put forward convincingly by Dąbrowska (2012, 2013), among others. She discusses a number of recent studies showing that adult monolingual native speakers of the same language do not share the same mental grammar. Dąbrowska argues that these differences may be caused by various factors. At times, it appears that speakers attend to different cues in the input. It may also well be the case that for certain constructions, some speakers extract only specific, 'local' generalizations, while others acquire more abstract rules. More educated speakers appear to acquire more general rules, possibly as a result of more varied linguistic experience.

There is reason to suspect that inter-individual variation may be particularly large when it comes to multi-word units. Language users are likely to share a large inventory of small, specific linguistic elements, such as single words and small chunks, e.g. HET BOEK, the choice of a neuter definite article in combination with the noun *boek*, as this combination is very frequent and alternatives, e.g. DE BOEK, the non-neuter definite article + *boek*, are (nearly) absent in the ambient language. Linguistic representations of larger, very general structures will be very similar too. An example of such a construction is the transitive pattern SUBJECT VERB OBJECT in which an Agent does something to a Patient. While the transitive sentences two speakers encounter will differ in content, the commonalities in meaning and structure enable the two speakers to arrive at similar abstract representations. People, most likely, differ to a larger extent in how, and how often, they encounter and use particular combinations of words and chunks. For example, the words *vast* (fixed, firm, certainly) and *zeker* (safe, certain, probably) are used frequently by both speakers of Belgian Dutch and speakers of Netherlandic Dutch. These two groups differ, however, in how they combine the two words in a multi-word unit that means 'definitely'. Both the orders *vast en zeker* and *zeker en vast* are observed. But Flemish speakers tend to prefer *zeker en vast* (at a ratio of approximately 4:1), whereas in the Netherlands *vast en zeker* is more frequent (at 7:1).[1] So, while Belgians and Dutch differ relatively little in usage frequency of the single words, they differ markedly in how and how often they use the two multi-word units and, presumably, in how familiar they consider each of them to be.

Investigations of the differences in language use between Belgians and Dutch are one example of the ways in which inter-individual variation is commonly studied: variation between speakers is examined by comparing groups that differ in terms of location (dialect), SES (sociolects) or ethnicity (ethnolects). However, also within such groups of speakers, there are likely to be differences between people in linguistic representations, as two persons are never identical in their language use and language exposure. In most linguistic judgment studies, variation between participants is either ignored, or reported as standard deviations but not discussed as a result in itself, or only taken into account by comparing groups of speakers. A usage-based perspective calls for an investigation that looks beyond such group averages. It also entails that differences between people in metalinguistic judgment are not sufficient to warrant the conclusion that these judgments are unreliable. Such differences may reflect genuine and meaningful differences in linguistic representations. In this study, the focus is on the variation, in order to shed a more complete light on the interplay of individual linguistic representations and the language system of a speech community.

In addition to *inter*-individual variation, a usage-based approach predicts *intra*-individual variation. If knowledge of a language in large part arises from usage, it is inherently dynamic. One's linguistic experiences change over time; one's linguistic representations are taken to change accordingly. Metalinguistic judgments based on changeable representations, therefore, are not expected to be stable over time. But what if the time frame is limited to a fairly short period in which the use of the word strings in question has not changed much? How (un)stable are people's judgments when they are to grade the same set of stimuli

[1] Ratios taken from the SoNaR corpus, a balanced, 500-million-word reference corpus of contemporary written Dutch texts of various styles, genres and sources, originating from the Dutch speaking area of Belgium (Flanders) and the Netherlands, as well as Dutch translations published in and targeted at this area (Oostdijk et al. 2013).

twice within a time span short enough for usage not to have changed much, yet long enough not to be able to recall the exact scores assigned the first time?

Even when usage frequency hasn't changed much for a particular stimulus, judgments regarding its familiarity may vary from one moment to the other due to differences in associations and the frame of reference used.[2] In judging familiarity, a speaker will activate potential uses of a given stimulus. The ease with which this is done, and the kinds of frames activated are highly dependent on the linguistic and extra-linguistic context. In the following section possible effects of context are discussed in more detail.

2.1.3 Context

Both the (extra-)linguistic context in which a participant encounters a stimulus and the (extra-) linguistic contexts the word string evokes, contribute to a frame of reference in which the stimulus is assessed. The extra-linguistic context — roughly speaking the setting in which the language use takes place— evokes scenarios a language user employs to interpret the linguistic input (Lakoff 1987), e.g. as a customer in a restaurant setting, it is perfectly fine to be told "let me tell you what today's specials are", followed by an enumeration of dishes. While clearly relevant for language use, this is not the type of context we focus on here. By having the participants in the current study perform the task in the exact same setting (location, experiment leader, instructions, format), we controlled for variation in the extra-linguistic context.

What we explore is how providing a sentential context for the stimuli may influence variation in metalinguistic judgments. Survey studies and studies of real-time language comprehension have shown that the immediate linguistic context affects the way in which word strings are interpreted, processed, and responded to (e.g. Camblin et al. 2007; Kamoen 2012). When it comes to empirical studies involving metalinguistic judgments, such context is usually deliberately absent. In lexical decision tasks, for example, the stimulus is the isolated word (or words) that participants must recognize, not a (non-)word in a sentence. For grammaticality judgments, the unit that is assessed is the isolated sentence (numerous examples in Sprouse et al. 2013). Any influence of linguistic elements

[2] One other obvious potential cause of intra-individual variation in familiarity ratings would be recent exposure, i.e. priming effects (e.g. Luka & Barsalou 2005; Schwanenflugel & Gaviska 2005). This is not the focus of the current study. Effects of recency, salience and other related concepts in exposure prior to a judgment task would have to be manipulated and/or measured systematically for participants. This would involve a tightly controlled experimental setting, with all linguistic exposure recorded.

other than the phenomenon under investigation, would usually be regarded as noise.

For judgments regarding the familiarity of units such as the prepositional phrases (PPs) we are investigating here, providing a context encapsulates the stimulus in a setting that makes it arguably more meaningful and realistic. In natural language use, these phrases do not occur in and of themselves; they occur in utterances. When a phrase is presented as an isolated word string, it may evoke different meanings and usage contexts across participants, and also within one person from one moment to another. Adding a context could reduce variation, as participants are prompted to focus on the same instance. For instance, when reading the words 'on the door', one may think of a poster hanging on the door, the practical joke with the bucket on the door, or someone knocking on the door. The number and kinds of usage contexts and the ease with which they come to mind will influence familiarity judgments. Diversity in associations may be related to differences in linguistic experiences, but it could also be more coincidental, resulting in less consensus among participants and more instability over time.

It is, as yet, an open question to what extent variation in familiarity judgments changes when the target items are embedded in a sentence. A sentential context activates a specific sense and generates an exemplar, which may guide the process of judging the item. For phrases that are used frequently, participants can easily come up with exemplars themselves. Presenting such frequent items in a sentence will probably not affect ratings much, provided that the sentence corresponds to participants' associations. Should the context not resemble the exemplars participants were thinking of, the scores may be lowered. For low-frequency stimuli, participants are more likely to have difficulties coming up with an exemplar. Giving a sentence context could then heighten the sense of familiarity, if it activates memory traces of very similar usage. If the given sentence context is not one that the participant recognizes, the effect could be that the item itself is rated as less familiar. Given that only one sense is mentioned, other possible uses of the item may not be taken into consideration. The PPs presented in this study were all fairly common phrases, many if not all of them polysemous or even homonyms (as [1]).

1. *Op de bank* *De jongens liggen op de bank televisie te kijken.*
on the couch/bank The boys lie on the couch television to watch
The boys are lying on the couch watching TV.

The context provided by the sentence in (1) is one that occurs frequently with this PP in the Corpus of Spoken Dutch, i.e. with an animate agent positioned [on the couch] involved in an activity. However, the word *bank* is a homonym; it can refer to a piece of furniture, as well as to a financial institution. The context generates a clear exemplar of the word in one sense, but at the same time rules out the other sense.

So, concluding: context may push the sense of familiarity up or down, depending on whether the provided context ties in with associations triggered in a participant's mind. Regardless of the direction, the expectation is that contexts reduce intra-individual variation in judgments as they steer what sense is evoked. Context may also reduce inter-individual variation as it stimulates participants to focus all on the same kind of exemplar, but this crucially depends on the extent to which a specific context is familiar to different participants. For high-frequency stimuli, effects of context are expected to be smaller. These stimuli are more likely to evoke the same kinds of exemplars across participants and at different points in time, and the contexts provided are likely to be recognizable to many of them.

2.1.4 Research questions

To start with, we examine the extent to which familiarity judgments are related to usage frequency and influenced by context. In our main analyses we investigate how stable these familiarity judgments are, looking at both inter- and intra-individual variation, and to what extent the stability varies depending on the frequency of the word combination and the presence of a context.

Given that familiarity ratings are taken to rest on usage frequency and similarity to other constructions, we expect to find a correlation between ratings and corpus frequencies. Furthermore, inter-individual variation in ratings is to be expected, since people differ in their linguistic experiences. Intra-individual variation is hypothesized to be smaller, as the rating sessions take place in a fairly short period in which the use of the word strings in question will not have changed much. We expect that embedding the stimuli in a context will reduce intra-individual variation in judgments as the context steers what sense is evoked. Whether or not context reduces inter-individual variation depends on the extent to which a specific context is familiar to different participants. Finally, the more frequent the item, the smaller effects of context are expected to be.

In other to test these hypotheses, we had participants judge the same linguistic stimuli twice within a relatively short period of time, in the same experimental setting. The data yield insight into the ways in which individual linguistic representations and the language system of a speech community are interrelated.

2.2 Method

2.2.1 Design

In order to test the stability of linguistic familiarity judgments for items with a range in frequency, and the influence of presenting these items in isolation or in context, a 2 (TIME) x 2 (CONTEXT) fully within-participant design was used. All participants rated 44 items both in isolation and in context, twice within the space of two to three weeks.

2.2.2 Participants

The participants were 86 students of Communication and Information Sciences at Tilburg University (66 female, 20 male) with an average age of 21.6 years. All of them were native speakers of Dutch. They participated for course credit.

2.2.3 Material

2.2.3.1 Stimulus items

Participants were asked to rate 44 Prepositional Phrases (PPs) consisting of a preposition and a singular noun, and in a majority of the cases a determiner (i.e. 35 with a definite article, 1 possessive *zijn* 'his'). An initial set of items was taken from V. Verhagen and Backus (2011) from which a selection was made based on two frequency characteristics: they represented a wide range in frequency (from 9 to 1066) in the approximately ten million word Corpus of Spoken Dutch (*Corpus Gesproken Nederlands,* henceforth CGN) and for all items this particular P–(Det)–N combination was the most frequent one compared to configurations with other determiners and inflectional forms of the noun (for a full list of items, and frequency data in CGN, see Appendices 2.1 and 2.2).[3]

For each PP a context sentence was created with a full lexical verb and often a nominal subject and object based on its occurrences in CGN (e.g. *in de kast* 'in the cupboard' often co-occurs with *leggen* 'lay', describing events in which someone puts something in a cupboard). The sentences were between 6 and 12 words long, with the PP occurring in the second half of the sentence but never as the final constituent, as in (2). We made sure not to refer to entities that may evoke strong feelings (e.g. 'Saddam Hussein'). All sentences are listed in Appendix 2.1.

2. *Ze heeft de spulletjes in de kast gelegd.*
   She has the little-stuff in the cupboard put.
   She put the things in the cupboard.

[3] CGN is a fairly small corpus. When SoNaR (a balanced reference corpus of contemporary written standard Dutch [Oostdijk et al. 2013]) became available, we investigated how often the items occur in the Netherlandic Dutch subset consisting of 143.83 million words. For both the PP as a whole and the noun (lemma search) there is a strong correlation between the CGN and the SoNaR frequencies (*r* = .93 and *r* = .90 respectively).

2.2.3.2 Judgment task

Participants were asked to rate familiarity using *Magnitude Estimation* (Bard et al. 1996). In this type of task, no set judgment scale is provided to the participants. Instead, participants rate each stimulus relative to the preceding one. This procedure requires a brief introduction and practice session (see Section 2.2.4). The construct of familiarity is clearly a gradual one, which fits well with the ratings provided by participants in a Magnitude Estimation task. Such a task allows participants to build their own scale. In contrast to a Likert scale, a Magnitude Estimation scale does impose a limited set of degrees of familiarity. The scale is open-ended, meaning that it is always possible to add higher or lower scores. Furthermore, participants are free to make as many fine-grained distinctions as they feel appropriate. Magnitude Estimation has been used successfully in judgments of grammatical well-formedness (e.g. Bader & Häussler 2010), productivity of morphological and modal verb constructions (Backus & Mos 2011) as well as idiomaticity (Wulff 2009). Among these, Wulff explicitly mentions that inter-subject consistency was extremely high, and Backus and Mos report high reliability measures (Cronbach's α = .85). In a follow-up study (reported on in Chapter 3), highly similar to the one reported here, we asked a new group of participants to give familiarity ratings at two points in time using either a Magnitude Estimation or a 7-point Likert scale. The type of scale does not appear to influence the degree of inter- and intra-individual variation much.

2.2.4 Procedure

The experiment was carried out in one computer room in the participants' faculty building under a research assistant's supervision. All participants completed the experiment twice, with a period of two to three weeks between the first and second session. They knew in advance that the experiment involved two test sessions, but not that they would be doing the exact same task twice. Given that the stimuli concern prepositional phrases that typically occur in everyday language use, our participants have about 20 years of linguistic experiences that contribute to their cognitive representations of these word strings. From that viewpoint, three weeks is a relatively short time span. Furthermore, there is no reason to assume that the use of the word combinations under investigation changes much in these three weeks. Therefore, the interval is not expected to bring about noticeable alterations in cognitive representations and metalinguistic judgments regarding the stimuli.

The items were presented in an online questionnaire form (using the Qualtrics software program) and this was also the environment within which the ratings were given. After signing a consent form and filling out a brief questionnaire regarding demographic variables (age, gender, language background), participants were introduced to the notion of relative ratings through the example of comparing the size of depicted clouds and expressing this relationship in numbers. They were instructed to rate each stimulus relative to the immediately preceding one, as this is what participants are inclined to do, rather than comparing each stimulus to a fixed modulus (e.g. Sprouse 2008). In a brief practice session, participants gave familiarity ratings to verb–object combinations (e.g. *veters strikken* 'to tie shoe laces'). Before starting the main experiment, they were given a few tips, i.e. not to restrict their ratings to the scale used in the Dutch grading system (1 to 10, with 10 being a perfect score), not to assign negative numbers, and not starting very low, to allow for subsequent lower ratings.

The main experiment consisted of two blocks: one in which the PPs were presented in isolation, and one with the PPs embedded in a sentence (with the PP underlined). Within each block, the order of presentation was randomized for each participant. Half of the participants started with the isolated block of items, the other half with the items in sentence contexts. The instructions were to rate familiarity of the word combination ("*Hoe vertrouwd vind je deze combinatie van woorden?*" – 'How familiar do you consider this combination of words?'). In earlier studies using familiarity ratings (e.g. Blasko and Connine 1999; Juhasz and Rayner 2003), the instructions for participants are very concise, illustrating that the term 'familiarity' can be understood without much introduction. Usually, participants are simply asked to rate how familiar they are with a stimulus on a 5- or 7-point Likert scale. When guidelines are provided, they refer to usage frequency. Williams and Morris (2004), for instance, asked participants to rate how often they had seen a given word. Juhasz et al. (2015, Appendix) used the phrasing "if you feel you know the meaning of the word and use it frequently, then give it a high rating on this scale".

Before judging the isolated word strings, our participants were told: If you wish, you could think of the combination in a particular context before judging it. Before rating the stimuli in sentences they were informed: You will see a word combination in a sentence. We would like to ask you to judge the familiarity of the underlined phrase in this specific context. We did not verify how carefully participants read the context. Given that the PP appeared in different positions on the screen, participants could not keep their eyes focused on one spot. The context consisted of just one sentence and it would have been difficult to refrain from reading it automatically.

2.2.5 Data transformations

For each participant, the ratings provided within one session were converted to Z-scores to make comparisons of relative ratings possible. This transformation is relatively common in acceptability judgments (Bader & Häussler 2010; Schütze & Sprouse 2013), as it involves no loss of information on ranking, nor at the interval level. By converting into Z-scores, a score of 0 indicates that a particular item is judged by a participant to be of average familiarity compared to the other items. For each item, Appendix 2.2 lists the mean of the Z-scores of all participants for that item, and the standard deviation.

To investigate the stability in judgment, a Z-score for an item in the second session was deducted from its score in the first session. The differences, or Δ-scores, were used to analyze the extent to which a participant rated an item differently over time (e.g. if a participant's rating for *naar huis* yielded a Z-score of 1.0 in the first session, and 0.5 in the second, the Δ-score is 0.5; if it was 1.0 the first time, and 1.5 the second time, the Δ-score is also 0.5, as the instability of the judgment is of the same magnitude). Absolute Δ-scores are used here, since it is of no importance for our research questions whether the difference in scores involves a higher or a lower score at Time 2. As participants constructed a scale at Time 1 and a new one at Time 2, ratings were converted into Z-scores at Time 1 and Time 2 separately. Consequently, we cannot determine whether participants might have considered all stimuli more familiar the second time. Since we used stimuli that are common in everyday language use, we have no reason to assume that their use and their perceived familiarity changed much within a period of two to three weeks. In order to investigate whether ratings move in one or another direction we need participants to use a fixed scale, for example a 7-point Likert scale. For this, we refer to the follow-up study in which a fixed scale was used (Chapter 3).

In order to relate familiarity judgments to frequency of the rated items, frequency counts of the exact word string in CGN were queried and subsequently log-transformed. The same was done for the frequency of the noun (lemma search). To give an example, the phrase *naar huis* occurred 1066 times in CGN, which corresponds to a log-transformed frequency score of 2.05. The lemma frequency of the noun, which encompasses occurrences of *huizen, huisje, huisjes* in addition to *huis*, amounts to 4730 instances. This corresponds to a log-transformed frequency score of 2.70. Figure 2.1 shows the positions of the stimuli on the phrase frequency scale and the lemma frequency scale; Appendix 2.2 lists for all stimuli the raw and the log-transformed frequencies.

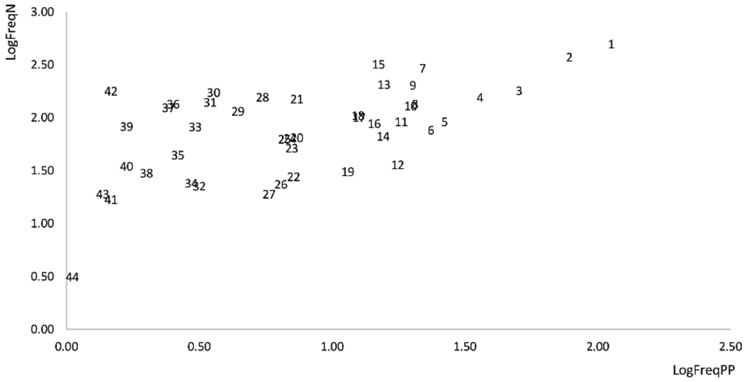


Figure 2.1 Scatterplot of the relationship between the log-transformed corpus frequency of the PP and that of the N (*r* = .59). The numbers 1 to 44 identify the individual stimuli (see Appendices).

2.2.6 Statistical analyses

First of all, we investigated to what extent the familiarity judgments can be predicted by the log-transformed frequency of the specific phrase (LogFreqPP) and the log-transformed lemma-frequency of the noun (LogFreqN), and to what degree the factors Context and Time (i.e. first or second session) exert influence. The stability of the judgments was investigated in a separate analysis.

We ran linear mixed-effects models (Baayen et al. 2008), using the function *lmer* from the lme4 package in the R software program (www.r-project.org). As Baayen and Milin (2010) state, mixed-models obviate the necessity of prior averaging over participants and/or items, and thereby offer the researcher the far more ambitious goal to model the individual response of a given participant to a given item.

In the first analysis, LogFreqPP, LogFreqN and Context were included as fixed effects, and so were all two-way interactions. Note that there cannot be a main effect of Time in this analysis, since scores were converted to Z-scores for the two sessions separately (i.e. the mean scores at Time 1 and Time 2 were 0). In the mixed-effects models we did include the two-way interactions of Time and the other factors. The fixed effects were standardized.

Participants and items were included as random effects. We incorporated a random intercept for items and random slopes for both items and participants to account for between-item and between-participant variation. The model does not contain a by-participant random intercept, because after the Z-score transformation all participants' scores have a mean of 0 and a standard deviation of 1. Furthermore, we excluded by-item random slopes for the factors LogFreqPP and LogFreqN, because each item has only one phrase frequency and one lemma frequency. Within these limits, a model with a full random effect structure was constructed following Barr et al. (2013). As the model did not converge, we excluded random slopes with the lowest variance step by step. When we obtained a converging model, a comparison with the intercept-only model proved that the inclusion of the by-item random slope for CONTEXT and the by-participant random slopes for the three fixed effects and for the interactions LOGFREQPP x CONTEXT and CONTEXT x TIME was justified by the data ($\chi^2(17) = 875.36$, $p < .001$).

In the second analysis, we investigated the stability of the judgments. We ran linear mixed-effects models on the Δ-scores computed for the ratings of each participant on each item in the two sessions (see Section 2.2.5 *Data transformations*). The absolute Δ-scores indicate the extent to which a participant's rating for a particular item at Time 2 differs from the rating at Time 1. For each item, we have a list of 86 Δ-scores that express each participant's stability in the grading. In order to fit a linear mixed-effects model on the set of Δ-scores, we log-transformed them using the natural logarithm function.[4]

We analyzed the log-transformed Δ-scores using linear mixed-models. LOGFREQPP, LOGFREQN and CONTEXT were included as fixed effects, participants and items as random effects. The fixed effects were standardized. We included a by-item random intercept and random slope for CONTEXT. For participants, we included a random intercept and random slopes for LOGFREQPP, LOGFREQN and CONTEXT. As the model did not converge, we excluded random slopes with the lowest variance step by step. When we obtained a converging model, a comparison with the intercept-only model proved that the inclusion of the by-subject random slopes for LOGFREQPP and CONTEXT was justified by the data ($\chi^2(5) = 79.28$, $p < .001$).

2.3 Results

2.3.1 Relating familiarity judgments to frequency and context

By means of linear mixed-effects models, we investigated to what extent the familiarity judgments can be predicted by the log-transformed frequency of the specific phrase (LOGFREQPP) and the log-transformed lemma-frequency of the noun (LOGFREQN), and to what degree the factors CONTEXT and TIME (i.e. first or second session) exert influence.[5] The resulting model is summarized in Table 2.1

[4] The absolute Δ-scores constitute the positive half of a normal distribution. Log-transforming the scores yields a normal distribution, thus complying with the assumptions of parametric statistical tests.

[5] Half of the participants first rated the phrases in isolation and then rated the same phrases embedded in a sentence; the other half did it the other way around. We tested whether this affected ratings by including the factor ORDERCONTEXT in

(confidence intervals were obtained via parametric bootstrapping over 100 iterations). The variance explained by this model is 33% (R^2m = .16, R^2c = .33).[6]

Table 2.1 Estimated coefficients, standard errors, and 95% confidence intervals for the mixed-model fitted to the familiarity ratings.

	B	*SE b*	*95 % CI*
Intercept	0.01	0.05	-0.09, 0.10
LogFreqPP	**0.46**	**0.07**	**0.34, 0.60**
LogFreqN	**-0.15**	**0.07**	**-0.27, -0.04**
Context	0.04	0.03	-0.01, 0.10
Context x LogFreqPP	-0.05	0.03	-0.10, 0.00
Context x LogFreqN	0.00	0.03	-0.04, 0.05
Context x Time	-0.02	0.01	-0.03, 0.00
LogFreqPP x Time	0.01	0.03	-0.01, 0.03
LogFreqN x Time	-0.01	0.01	-0.03, 0.01
LogFreqPP x LogFreqN	-0.01	0.04	-0.10, 0.07

Note. Significant effects are printed in bold.

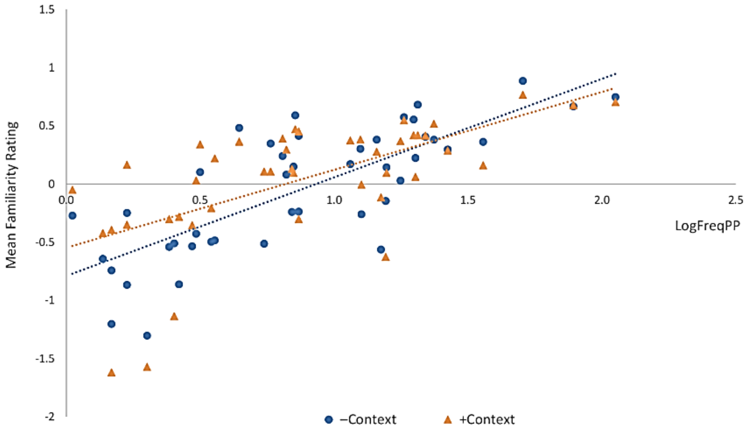


Figure 2.2 Scatterplot of the log-transformed corpus frequency of the PP and its mean familiarity rating.

the mixed-effect models. The order of the Context-block and the No Context-block did not have a significant effect on judgments (B = 0.00; SE = 0.01; 95% CI = -0.01, 0.01).

[6] R^2m (Marginal R_GLMM²) represents the variance explained by fixed effects; R^2c (Conditional R_GLMM²) is interpreted as variance explained by both fixed and random effects (i.e. the entire model).

First of all, the model shows an effect of LogFreqPP. Log-transformed frequency of the phrase in CGN significantly predicted judgments, with higher frequency leading to higher familiarity ratings, as can be observed from Figure 2.2.

Figure 2.2 also shows certain differences between items that were presented in a sentence (orange triangles) and items that were presented as isolated word strings (blue dots). For *low*-frequency phrases, providing a context tended to heighten the ratings; in the *middle* part of the frequency range, there is very little difference between +Context and –Context items; and for *high*-frequency phrases, adding a context slightly lowered the ratings. However, these differences were not pronounced enough for the interaction between Context and LogFreqPP to be significant (note that the confidence interval for the Context x LogFreqPP interaction is [-0.10, 0.00]).

A factor that did prove to have a significant effect is LogFreqN. Higher frequency of the noun resulted in *lower* familiarity ratings for the prepositional phrase. While significant, this effect was not as strong as that of phrase frequency. Figure 2.3 shows the mean familiarity ratings in relation to the log-transformed frequency of the noun. Note that higher noun frequency often entails higher phrase frequency. While the former results in lower ratings, the latter leads to higher ratings. Since phrase frequency has a stronger effect than noun frequency, one cannot observe a clear descending line in Figure 2.3.

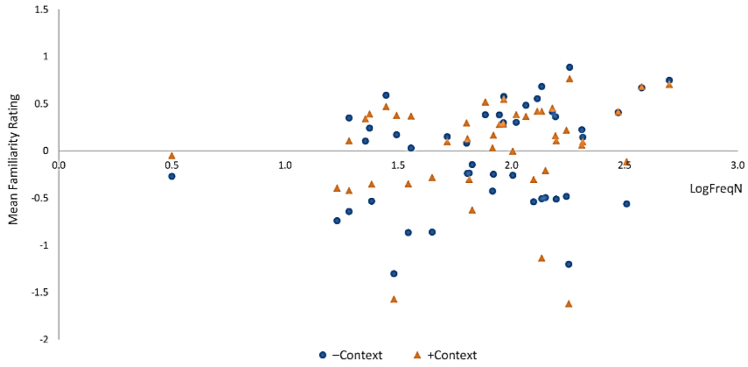


Figure 2.3 Scatterplot of the log-transformed corpus frequency of the N and the mean familiarity rating of the PP as a whole.

2.3.2 Stability of familiarity ratings

To examine the stability of the familiarity judgments, we calculated the correlation between the ratings assigned at Time 1 and those assigned at Time 2. When averaging over participants, the ratings are highly stable. Mean ratings were computed for each of the 88 items at Time 1, and likewise at Time 2. The correlation between these two sets of mean ratings is nearly perfect (Pearson's r = .97).

Comparisons of *individual participants'* ratings at Time 1 and Time 2 show a rather different picture. For each participant we computed the correlation between that person's judgments at Time 1 and that person's judgments at Time 2. This yielded 86 correlation scores that range from -.13 to .87, with a mean correlation of .52 (SD = .20). This means that none of the participants is as stable in their ratings as the aggregated ratings are, and some participants (N = 5 with correlations < .10) show very little if any correlation with their own ratings, i.e. their ratings at Time 2 do not correlate at all with the ratings on the same items, with the same instructions and under the same circumstances a few weeks earlier.[7] Two-thirds of the participants had self-correlation scores between .32 and .70. Figure 2.4 shows the distribution of the correlations of our 86 participants.

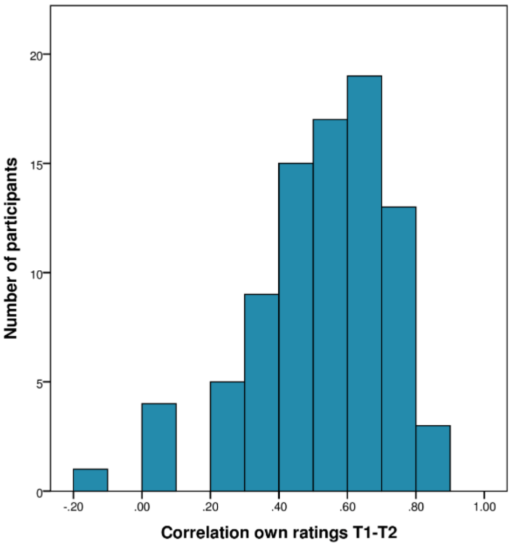


Figure 2.4 Distribution of participants' correlation of their own ratings (Pearson's r, Time 1 – Time 2).

[7] These five participants with T1-T2 correlations <.10 stand out. We identified these participants and examined their judgments in more detail. This is discussed in relation to Figure 2.7 below.

If there are stable individual differences, participants' ratings at Time 1 should be more similar to their own ratings at Time 2 than to the other participants' ratings at Time 2.[8] We compared each participant's self-correlation to the correlation between that person's ratings at Time 1 and the group mean at Time 2 by means of the procedure described by Field (2013: 287). For 16 participants, self-correlation was significantly higher than correlation with the group mean; for 17 participants correlation with the group mean was significantly higher than self-correlation; for 53 participants there was no significant difference between the two measures.

In order to determine if familiarity ratings were stable for *certain items* more so than for others, we used the Δ-scores (see Section 2.2.5 *Data transformations* and Section 2.2.6 *Statistical analyses*). Figure 2.5 shows for each item the mean log-transformed Δ-score. The lower this Δ-score, the more stable the judgments were. A Δ-score of 0.02 (meaning very little difference between the ratings at Time 1 and Time 2) corresponds to a log-transformed Δ-score of -3.91. As can be observed from Figure 2.5, none of the items approaches the value -3. This indicates that none of the items elicited stable ratings from *all* participants.

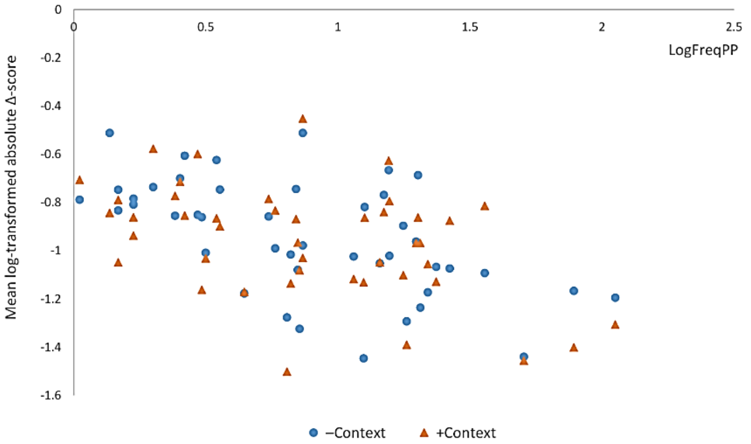


Figure 2.5 Scatterplot of the log-transformed corpus frequency of the PP and its mean log-transformed absolute Δ-score.

We analyzed the log-transformed Δ-scores using linear mixed-models. The resulting model is summarized in Table 2.2 (confidence intervals were obtained via parametric bootstrapping over 100 iterations). Only LogFreqPP proved to have

[8] I would like to thank an anonymous reviewer for this suggestion.

a significant effect. Higher phrase frequency led to less instability in judgment. The variance explained by this model is 14% (R2m = .01, R2c = .14). In comparison to the relation between frequency and judgment, the relation between frequency and instability is less strong.

Table 2.2 Estimated coefficients, standard errors, and 95% confidence intervals for the mixed-model fitted to the log-transformed absolute Δ-scores.

	b	*SE b*	*95 % CI*
Intercept	-0.95	0.05	-1.07, -0.84
LogFreqPP	**-0.14**	**0.04**	**-0.21, -0.07**
LogFreqN	0.04	0.04	-0.03, 0.11
Context	-0.01	0.03	-0.05, 0.03
Context x LogFreqPP	0.02	0.02	-0.01, 0.06
Context x LogFreqN	-0.01	0.02	-0.05, 0.03
LogFreqPP x LogFreqN	0.00	0.03	-0.04, 0.05

Note. Significant effects are printed in bold.

In sum, both phrase frequency and noun frequency proved significant predictors of familiarity judgments. Embedding the phrases in a sentence did not have a significant effect on the familiarity ratings. Regarding the stability of judgments we observed that, as a group, the participants provide a very stable pattern of familiarity ratings: the overall rankings at Time 1 and Time 2 correlate nearly perfectly. As soon as one zooms in on individual participants, or looks at individual items, the picture becomes less stable.

2.4 Discussion

2.4.1 Coexisting stability and instability

Our study reveals that stability is found in the average judgments; individuals' judgments display much more variability. The picture that arises from our data is one of two perspectives. On the one hand, there is a particularly strong correlation between the average ratings on Time 1 and on Time 2, as well as a clear correlation between those average familiarity ratings and log-transformed corpus frequencies of the word strings. This is where the stability resides. On the other hand, a large majority of the participants provided rather different ratings at Time 2 compared to Time 1; none of the participants was as stable in their ratings as the aggregated ratings are; and no single item elicited stable ratings from all of our participants. There is clear variation in individual ratings. While these results may seem to be at odds, we feel that they provide a very real portrayal of metalinguistic representations and, possibly, linguistic representations.

One way to look at our dataset is to think of it as two photo mosaics created at different times. Each mosaic is composed of numerous little photographs – in our case: the 7568 ratings we collected within one session (86 participants each rated 88 items). When you zoom out, all these different elements together yield one picture. By having the participants rate the stimuli a second time, we obtained a second picture. From a distance, these two pictures look very similar; as you zoom in you will notice differences. Within one picture, you see that any given part is composed of multiple elements that differ from each other to a greater or lesser extent (i.e. the various degrees of inter-individual variation, described in the double-framed box in Figure 2.6). Furthermore, when you compare the two pictures in detail, you will find that a given individual element is not exactly the same at T1 and T2 (i.e. the various degrees of intra-individual variation, described in the circle in Figure 2.6).

The similarity between the two pictures that is observed when you zoom out (i.e. the stability of the average ratings visible in the near perfect T1-T2 correlation) ties in well with the idea that the language system of a speech community appears to be quite robust. In order to ensure intelligibility and learnability of the language, this system must not change too much –especially in the short space of a couple of weeks, and concerning everyday linguistic units like the prepositional phrases tested in this study.

However, for the overall picture to be stable, it is not necessary for all of the component parts to be constants too. This is shown in the fact that no single participant's ratings proved to be as stable as the average ratings, and the fact that no single item elicited stable ratings from *all* participants. The fact that the *overall* ratings correlated perfectly means that as some participants gave a higher rating the second time, others gave a lower rating, such that on average an item's score remains the same. The individual variation does not entail overall instability.

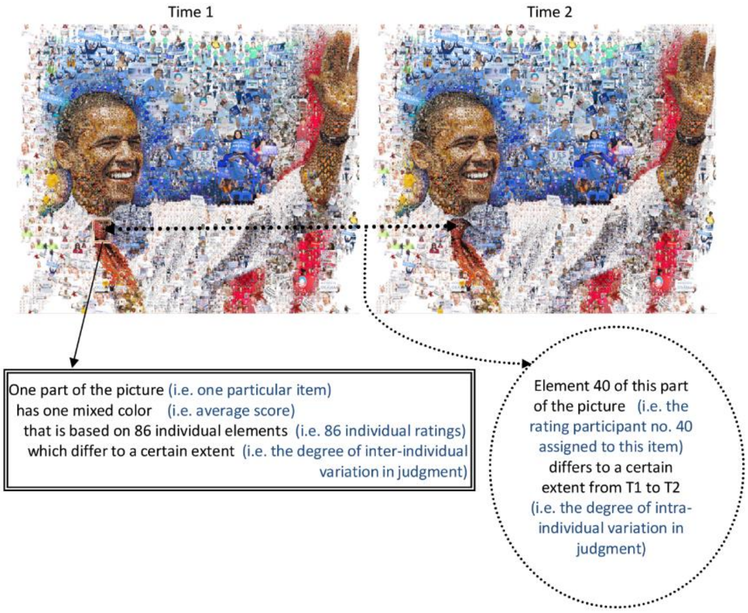

Time 1
Time 2
One part of the picture (i.e. one particular item)
has one mixed color (i.e. average score)
that is based on 86 individual elements (i.e. 86 individual ratings)
which differ to a certain extent (i.e. the degree of inter-individual variation in judgment)
Element 40 of this part of the picture (i.e. the rating participant no. 40 assigned to this item)
differs to a certain extent from T1 to T2
(i.e. the degree of intra-individual variation in judgment)

Figure 2.6 Visualization of inter- and intra-individual variation by means of two photo mosaics composed of numerous little photographs. Adapted from *Hope over Fear* (2008) by C. Tsevis. Copyright holder unknown. Retrieved from http://www.dripbook.com/tsevis/illustration-portfolio/barack-obamai/#288337. Adapted with permission.

2.4.2 Sources of inter- and intra-individual variation

If the overall picture is remarkably stable, what does the inter- and intra-individual variation tell us? In the following paragraphs we explore possible causes for the intra-individual variation from Time 1 to Time 2. One possible cause for a change in familiarity score is a change in use: some phrases may become more familiar, others less so. However, it is highly improbable that our participants' use of the prepositional phrases in question changed a lot in the course of a few weeks, let alone that some PPs became more frequent for certain participants and less so for others such that on average the items' scores remained the same.

The observed variation could be noise inherent in the process of judging. Featherston (2007) contends that each individual judgment is noisy and that most of the differences between individuals are just error variance. Mean judgments effectively remove this error variance, since -random- errors cancel each other out. If you test groups, Featherston argues, you will see that groups of respondents agree quite closely. The latter is borne out, as there is a near perfect correlation between the average ratings on Time 1 and on Time 2. It is not self-evident, though, that all differences between individual judgments can be considered noise. True noise is fairly random fluctuation around the group pattern. Why then were some participants' judgments remarkably stable (i.e. *r* = .87 and *r* = .80 in Figure 2.4)? And why did we observe significantly less instability for high-frequency phrases compared to lower-frequency items (Table 2)? There seems to be more to it than just random variance. It would be interesting to determine how much of the variation could be considered noise. One way to examine this in future research would be to assume an identical grammar and simulate different amounts of noise, and then compare the results with the experimental data.

One factor that may have increased noise is task-related: some of the instability over time may well be due to the rating scale used. According to some researchers (e.g. Weskott & Fanselow 2011), Magnitude Estimation is more likely to produce variance than Likert scale or binary judgment tasks, due to the increased number of response options in ME. However, several other studies (e.g. Bard et al. 1996; Wulff 2009; Bader & Häussler 2010) provide evidence that ME yields reliable data, not different from those of other judgments tasks, and that inter-participant consistency is extremely high. Leaving it to participants to construct their own response scale is a considerable advantage of ME in judgment tasks where the construct of interest is gradient. From a usage-based perspective of language, this is the case for most metalinguistic judgments and especially so for familiarity ratings. It could be argued, moreover, that this self-construal of a rating scale involves deeper, more considered processing and evaluation of the stimulus items. If anything, that would predict stronger memory traces and therefore a higher correspondence in ratings between Time 1 and Time 2. A follow-up study (Chapter 3), in which the use of Magnitude Estimation and a 7-point Likert scale was compared, shows slightly less intra-individual variation for participants using ME than for those using Likert scale ratings.

In addition to the unrestricted number of response options available to a participant, another characteristic of ME might play a role. In Magnitude Estimation, a rating for an item is given in comparison to a previous item. In our data collection, the order of items was randomized automatically for each participant both at Time 1 and at Time 2. The simple fact that a participant rated a particular item A after item B at Time 1, but after item C at Time 2 will have influenced the rating of item A (see also Sprouse 2011). Since the software program did not record the orders in which the stimuli were presented, we cannot determine how much of the variance is explained by the rating a participant assigned to the previous stimulus. However, the presentation order is unlikely to account for the fact that there were very large differences between participants in the stability of their ratings (see Figure 2.4). As the experiment consisted of two sets of 44 items that were randomized, it is improbable that for certain participants the orders at T1 and T2 were nearly identical.

Another possible source for the inter- and intra-individual variation in familiarity ratings is that participants performed the task using different strategies. It is unlikely that respondents were simply not paying attention. The co-presence of a research assistant encouraged them to carry out the task attentively. What is more, the clear correlation between ratings and corpus frequencies and the extremely stable mean ratings are not to be expected if they had not performed the task seriously. What may be the case is that participants took different things into account while giving scores. Hintzman (2011) claims that relationships of repetition, exposure duration, and recency are falsely reduced to a single underlying process: familiarity. His critique mainly concerns recognition-memory and free recall paradigms that test people's ability to indicate what words were in a given list. In our study, repetition, exposure duration, and recency within the experimental setting were practically the same for all participants. However, one could suspect that people who provided relatively stable judgments based their ratings on the same considerations at Time 1 and Time 2, whereas the participants whose judgments were unstable took into account various aspects, including ones that differed from Time 1 to Time 2. One way to explore this, is to investigate whether the stable judgments correlate better with corpus frequencies than the unstable judgments do.

To see whether stable people base their ratings on frequency, we plotted each participant's stability (ranging from -1: absolute negative correlation between one's ratings at Time 1 and Time 2, to +1: absolute positive correlation between one's ratings at Time 1 and Time 2) against the extent to which his/her scores reflect corpus frequencies. Each circle in Figure 2.7 represents a participant. There are five clear outliers when it comes to participants' correlations with their own ratings (x-axis). These five participants correlate less than .10. The correlations between their scores and the corpus frequencies (y-axis) range from -.05 to .48. For the other 81 participants, this correlation ranges between .21 to .66. Upon closer inspection, only the participant with a negative self-correlation and a negative rating-corpus frequency correlation is a true outlier. Furthermore, the cluster of 81 participants still shows a considerable range on both axes. While there is a relationship between the two measures (*r* = .29), we observe substantial dispersal. Participants whose scores do not correlate with corpus frequencies can still be pretty consistent and, conversely, those whose judgments are correlated with frequency, are not necessarily consistent in their ratings.

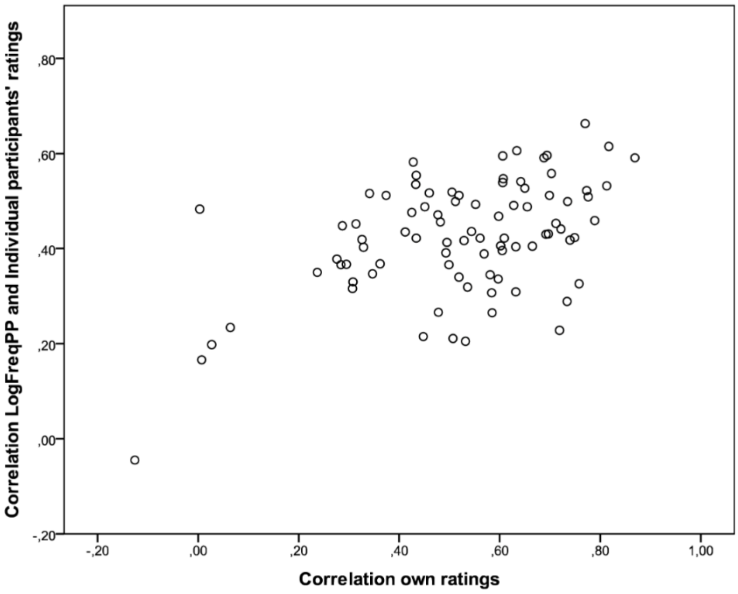


Figure 2.7 Scatterplot of participants' stability in their own ratings (Pearson's *r*, Time 1 – Time 2) against the correlation between their ratings and the log-transformed frequencies of the PPs (Pearson's *r*). The extent to which a participant's position on the x-axis is correlated with the position on the y-axis is *r* = .49 when all participants are included, and *r* = .29 when the five participants whose own T1-T2 correlation is less than .20 are excluded.

In sum, although there is noise contained in a single judgment, random noise does not seem to account for the patterns of variation in the data. Therefore, we would like to explore the possibility that variation may be a genuine property of one's metalinguistic representations and ultimately one's linguistic representations.

2.4.3 Reconsidering the nature of metalinguistic judgments

Crucially, the observed *intra*-individual variation within a short period of time prompts us to reconsider interpretations of *inter*-individual variation. The intra-individual variation shows that for quite a few of our participants metalinguistic judgments are not particularly stable, and thus suggests that also one's (meta)linguistic representations vary. Thus, at Time 1 participant Y may assign a higher score to a particular item than participant Z does, while at Time 2 it is the other way around. If this intra-individual variation over time reflects the genuine dynamism of linguistic representations, the difference between participant Y's rating and participant Z's rating at one point in time cannot be interpreted straightforwardly as *the* difference in their linguistic representations. A more complete and more faithful impression requires multiple measurements.

Our data show a clear correlation between mean familiarity ratings and corpus frequencies of the stimuli — both types of scores being amalgamations of data of different people. The log-transformed frequency of the PP as well as the noun was found to be a significant predictor of familiarity ratings. As hypothesized, higher phrase frequency led to higher ratings. Higher frequency of the noun resulted in lower ratings. The more frequent the noun, the more likely it is that this noun also occurs in other phrases that are frequently used.[9] Such phrases may come to mind when rating the stimulus. If some of them are considered more familiar, the score assigned to the stimulus is likely to be lowered.

Interestingly, phrase frequency had an effect –albeit small– on the degree of *intra*-individual variation in judgments. Higher phrase frequency led to more stability in judgment (see Figure 2.2). As for the degree of *inter*-individual variation, low-frequency phrases were found to display more variation in judgment across participants (as evidenced by higher SDs). There are several ways to interpret these findings. In all likelihood the use of items that are infrequent in the corpus differs more across our participants than the use of higher-frequency items does. An item with a low corpus frequency may be fairly common for some people, while others virtually never use it. As a result, familiarity judgments for this item will diverge. It could also be the case that even when actual usage frequency is comparable across participants, low-frequency items tend to yield more variation in judgments because people differ in the number and type of associations and exemplars that become activated much more so than for higher-frequency items.

A question awaiting further research is to what extent the degrees of inter- and intra-individual variation vary as the design of the experimental task changes (e.g. different instructions, other types of stimuli, or a task that measures immediate

[9] This is substantiated by corpus frequencies taken from the Netherlandic Dutch subset of SoNaR. The log-transformed frequency per million words of the phrase *in het water* is 1.08. *Water* is a high-frequency noun (logN 2.30) that occurs in various prepositional phrases, e.g. *op het water* (logPP 0.47), *onder water* (logPP 0.83), *boven water* (logPP 0.82), *uit het water* (logPP 0.38). A similar pattern is observed with respect to *in de hand* (logPP 1.34). *Hand* (logN 2.70) occurs in various prepositional phrases, e.g. *aan de hand* (logPP 1.79), *voor de hand* (logPP 1.37), *uit de hand* (logPP 0.38). The noun *bad*, by contrast, occurs less often (logN 1.56). When used together with a preposition, the phrase *in bad* is the most frequent combination (logPP 0.89). Other phrases are much less frequent: *uit bad* (logPP -0.30), *met bad* (logPP -1.08).

language processing). Possibly, individual participants' ratings are more stable if the stimuli cover a larger frequency range, as the differences between the stimuli are clearer to them. It would also be interesting to investigate the effects of including (a) phrases that don't occur much but plausibly could occur and (b) phrases that don't occur and require some thought to make sense of (cf. Caldwell-Harris et al. 2012). Recall that familiarity of a word sequence is taken to rest on frequency and similarity to other words, constructions or phrases. We would therefore predict that phrases of type (a) are more likely to resemble phrases that are familiar, and would receive higher ratings than phrases of type (b). Furthermore, it remains to be seen what patterns of inter- and intra-individual variation emerge when items other than multi-word phrases are investigated. As we suggested before, multi-word sequences may involve more variation in usage and perceived familiarity than single words and highly schematic constructions. To the extent that tasks differ in the processes and knowledge they tap into, the explanatory power of specific corpus-based measures may vary (see Wiechmann 2008 and Divjak 2016 for insightful comparisons). While phrase and noun frequency were shown to have explanatory power for the familiarity ratings, there is variance they could not explain, meaning that there are other variables influencing these ratings.

In our study, we examined the factor CONTEXT, as we expected familiarity judgments to be influenced by the number and type of usage contexts that come to mind. More specifically, we hypothesized that the presence of a sentential context generates an exemplar which may affect both the familiarity ratings and the degrees of inter- and intra-individual variation. While the factor CONTEXT did not yield a significant effect, the observed trend is in the expected direction. Most items at the lower end of our frequency scale were rated higher when presented with rather than without a context (see Figure 2.2). It is likely that participants have more difficulty to come up with exemplars for low- than for higher-frequency phrases. When the phrase is embedded in a sentence, the participant is offered an exemplar of the item in use. If this exemplar is considered recognizable —that is, if it activates memory traces of very similar usage— it will heighten the sense of familiarity. In the mid- and high-frequency range, there is very little difference between the ratings assigned to +Context and –Context items. Given that participants have more experiences with higher-frequency phrases, it will be easier for them to think of exemplars. As we strove to formulate prototypical contexts, the given sentence is likely to resemble the exemplars participants were thinking of. In those cases, adding a context does not alter judgments. With a view to the comparability of the results of different judgment tasks and their meaningfulness and generalizability, it could be considered reassuring that the presence or absence of a sentential context does not appear to yield significantly different judgments.

The fact that providing a context did not influence the degrees of inter- and intra-individual variation is puzzling though. We predicted that adding a context would result in less variation in judgments, as the context steers what sense is evoked. This was not borne out by the data: making participants focus on the same kind of usage context did not systematically reduce variation in judgment across participants or over time. To adequately explain this observation requires a more elaborate investigation of the factor Context. Possibly, participants differ in how appropriate or prototypical they considered a given context to be. This may be related to differences in their linguistic experiences, or it may involve other factors. It would be interesting to explore effects of context in more detail by using larger text fragments, systematically varying their prototypicality, and using a think-aloud protocol to gain insight into participants' associations and considerations.

2.4.4 Conclusion

This article started by pointing out that judgments are often used as sources in linguistic research, while, really, there is much that we do not know regarding the reliability of these judgments. For now, our findings encourage us to continue using judgment data, but to see these data in a slightly different light. These judgment scores can be investigated at various levels of granularity. Aggregated means reflect perhaps not so much an idealized speaker/hearer in the Chomskian sense, but the overall stability of the language system in a speech community. The individual variation can be revealing as well; variation and instability is likely to be a genuine characteristic of (meta)linguistic representations.

We recommend, first and foremost, that researchers reflect on and be explicit about their object of interest: mental representations that speakers have and/or the systematicities and patterns in a language as spoken in a speech community. Depending on one's research focus, confining oneself to average scores or a single measurement may result in an incomplete and oversimplified picture. Average scores and corpus frequencies are only adequate if your unit of analysis is at the community level. Crucially, the cognitive linguistic framework urges researchers to expand their investigations beyond this level. Given that people's representations of language are mental entities, cognitive linguists cannot restrict themselves to aggregated data. Furthermore, from a usage-based perspective, both inter- and intra-individual variation are core characteristics of language.

If cognitive linguists take their usage-based principles seriously, they ought to pay more attention to variation both in their research design and in the analysis and interpretation of their data. Therefore, multiple measurements should become the norm rather than the exception. They are necessary in order to get a reliable picture of the dynamism of linguistic representations. This is in keeping with observations that the activation and processing of linguistic units can vary from one moment to the other. As Dąbrowska (2014: 646) puts it: "(…) even the same speaker may assemble the same utterance using different chunks on different occasions, depending on, for example, which units have been primed by prior discourse. This flexibility helps to explain the speed of language processing: we save time by opportunistically using whichever chunks are most accessible at the time of the speech event."

Importantly, variation may be more than something that is caused by communicative demands and affordances. It may be more than 'noise' in *performance* disturbing our view of *competence*. Perhaps the underlying linguistic representations, too, are more variable than commonly assumed. While the dynamic nature of representations lies at the heart of usage-based approaches, it is as yet not clear how much variability is to be expected within different time frames, for specific and more schematic units. A better understanding of patterns of variation will contribute to a more adequate model of linguistic representations. At the moment, we cannot be certain to what extent the variability of metalinguistic judgments reflects the variability of linguistic representations. Still, even if factors such as priming and salience are the causes of the observed variability, that momentary linguistic experience (i.e. the language that is produced, perceived and/or judged) is taken to exert some influence on one's representations. How strongly and directly prior and recent experiences influence linguistic representations, and how precisely metalinguistic judgments and processing measures reflect these representations, are questions awaiting further research.

Chapter 3

Abstract

In a usage-based framework, variation is part and parcel of our linguistic experiences, and therefore also of our mental representations of language. In this paper, we bring attention to variation as a source of information. Instead of discarding variation as mere noise, we examine what it can reveal about the representation and use of linguistic knowledge. By means of metalinguistic judgment data, we demonstrate how to quantify and interpret four types of variation: variation across items, participants, time, and methods. The data concern familiarity ratings assigned by 91 native speakers of Dutch to 79 Dutch prepositional phrases such as in de tuin 'in the garden' and rond de ingang 'around the entrance'. Participants performed the judgment task twice within a period of one to two weeks, using either a 7-point Likert scale or a Magnitude Estimation scale. We explicate the principles according to which the different types of variation can be considered information about mental representation, and we show how they can be used to test hypotheses regarding linguistic representations.

This chapter is based on:

Verhagen, V., Mos, M., Schilperoord, J., & Backus, A. (2019). Variation is information: Analyses of variation across items, participants, time, and methods in metalinguistic judgment data. *Linguistics*. Advance online publication. https://doi.org/10.1515/ling-2018-0036

Acknowledgements

I thank Carleen Baas for her help in collecting the data and Martijn Goudbeek for his helpful comments on the paper we submitted to Linguistics.

Chapter 3 Variation is information: Analyses of variation across items, participants, time, and methods in metalinguistic judgment data

3.1 Introduction

The past decades have witnessed what has been called a quantitative turn in linguistics (Gries 2014, 2015; Janda 2013). The increased availability of big corpora, and tools and techniques to analyze these datasets, gave major impetus to this development. In psycholinguistics, more attention is being paid to the practice of performing power analyses in order to establish appropriate sample sizes, reporting confidence intervals, and using mixed-effects models to simultaneously model crossed participant and item effects (Cumming 2014; Baayen et al. 2008; Maxwell et al. 2008). In research involving metalinguistic judgments great changes occurred. As Schütze and Sprouse (2013: 30) remark, "the majority of judgment collection that has been carried out by linguists over the past 50 years has been quite informal by the standards of experimental cognitive science". Theorizing was commonly based on the relatively unsystematic analysis of judgments by few speakers (often the researchers themselves) on relatively few tokens of the structures of interest, expressed by means of a few response categories (e.g. "acceptable", "unacceptable", and sometimes "marginal"). This practice has been criticized on various accounts (e.g. Dąbrowska 2010; Featherston 2007; Gibson & Federenko 2010, 2013; Wasow & Arnold 2005), which led to inquiries involving larger sets of stimuli, larger numbers of participants, and/or multiple test sessions. An unavoidable consequence is that the range of variation that is measured increases tremendously. Whenever research involves multiple measurements, there is bound to be variation in the data that cannot be accounted for by the independent variables. Various stimuli instantiating one underlying structure might receive different ratings; different people may judge the same item differently; a single informant might respond differently when judging the same stimulus twice. A question that then requires attention is: what to make of the variability that is observed? In this paper, we attempt to strike a balance between variation that is 'noise' and variation that is information, and we attempt to lay out the principles underlying this balance. Four types of variation will be discussed: variation across items, variation across participants, variation across time, and variation across assessment methods. We will explicate the principles according to which these types of variation can be considered informative, and we will show how to investigate this by means of a metalinguistic judgment task and corpus data.

First of all, there may be variation across items that are intended to measure the same construct (see Cronbach 1951 on Cronbach's alpha, H. Clark 1973 on the language-as-fixed-effect fallacy, and Baker & Seock-Ho 2004 on Item Response Theory and the Rasch model). If these stimuli yield different outcomes, this could lead to a better understanding of the influence of factors other than the independent variables under investigation. For example, acceptability judgments may appear to be affected by lexical properties in addition to syntactic ones. More and more researchers realize the importance of including multiple stimuli to examine a particular construct and inspecting any possible variation across these items (e.g. Featherston 2007; Gibson & Federenko 2010, 2013; Wasow & Arnold 2005).

Secondly, when an item is tested with different participants, hardly ever will they all respond in exactly the same manner. While it has become fairly common to collect data from a group of participants, there is no consensus on what variation across participants signifies. The way this type of variation is approached and the extent to which it plays a role in research questions and analyses depends, first and foremost, on the researcher's theoretical stance.

If one assumes, as generative linguists do, that all adult native speakers converge on the same grammar (e.g. Crain & Lillo-Martin 1999: 9; Seidenberg 1997: 1600), and it is this grammar that one aims to describe, then individual differences are to be left out of consideration. An important distinction, in this context, is that between competence and performance. Whenever the goal is to define linguistic competence, this competence can only be inferred from performance. When people apply their linguistic knowledge –be it in spontaneous language use or in an experimental setting– this is a process that is affected by memory limitations, distractions, slips of the tongue and ear, etc. As a result, we observe variation in performance. In this view, variation is caused by extraneous factors, other than competence, and therefore it is not considered to be of interest. In Chomsky's (1965: 3) words: "Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance."

Featherston (2007), a proponent of this view, explicitly states that variation in judgment data is noise inherent in the process of judging. Consequently, one should not compare individuals' judgments. As he puts it: "each individual brings their own noise to the comparison, and their variance in each judgement may be in opposite directions” (pp.284-285). As a result, individuals’ judgments seem to differ considerably, while most of the difference is just error variance. Featherston’s advice is to collect judgments from different participants and to average these ratings. In this way, “the errors cancel each other out and the judgements cluster around a mean, which we can take to be the ‘underlying’ value, free of the noise factor” (p.284).

A rather different approach to variation between speakers can be observed in sociolinguistics and in usage-based theories of language processing and representation. In these frameworks, variation is seen as meaningful and theoretically relevant. Characteristic of sociolinguistics is “the recognition that much variability is structured rather than random” (Foulkes 2006: 649). Whereas Featherston argues that variation is noise, Foulkes (2006: 654) makes a case for variability not to be seen as a nuisance but as a universal and functional design feature of language. Three waves of variation studies in sociolinguistics have contributed to this viewpoint (Eckert 2012). In the first wave, launched by Labov (1966), large-scale survey studies revealed correlations between linguistic variables (e.g. the realizations of a certain phoneme, the use of a particular word) and macro-sociological categories of socioeconomic class, sex, ethnicity, and age. The second wave employed ethnographic methods to explore the local categories and configurations that constitute these broader categories. The third wave zooms in on individual speakers in particular contexts to gain insight into the ways variation is used to construct social meaning. It is characterized by a move from the study of structure to the study of practice, which tends to involve a qualitative rather than quantitative approach.

A question high on the agenda is how these strands of knowledge about variability can be unified in a theoretical framework (Foulkes 2006: 654). Usage-based approaches to language processing and cognitive linguistic representations show great promise. As Backus (2013: 23) remarks: “a usage-based approach (…) can provide sociolinguistics with a model of the cognitive organization of language that is much more in line with its central concerns (variation and change) than the long-dominant generative approach was (cf. Kristiansen & Dirven 2008).”

From a usage-based perspective, variation across speakers in linguistic representations and language processing is to be expected on theoretical grounds. In contrast to generative linguistics, usage-based theories hold that competence cannot be isolated from performance; competence is dynamic and inextricably bound up with usage. Our linguistic representations are form-meaning pairings that are taken to emerge from our experience with language together with general cognitive skills and processes such as schematization, categorization and chunking (Barlow & Kemmer 2000; Bybee 2006; Tomasello 2003). The more frequently we encounter and use a particular linguistic unit, the more it becomes entrenched. As a result, it can be activated and processed more quickly, which, in turn, increases the probability that we use this form when we want to express the given message, making this construction even more entrenched. Language processing is, thus, to a large extent driven by our accumulated linguistic experiences, and each usage event adds to our mental representations, to a larger or lesser extent depending on its salience.[10]

Given that people differ in their linguistic experiences, individual differences in (meta)linguistic knowledge and processing are to be expected on this account. Such variation is arguably less prominent at the level of syntactic patterns compared to lexically specific constructions. Even though people differ in the specific instances of a schematic construction they encounter and use, they can arrive at comparable schematic representations. Still, even in adult native speakers' knowledge of the passive, a core construction of English grammar, individual differences have been observed (Street & Dąbrowska 2014).

The role of frequency in the construction and use of linguistic representations in usage-based theories has sparked interest in variation across speakers. Various studies (Balota et al. 2004; Caldwell-Harris et al. 2012; Dąbrowska 2008; Street & Dąbrowska 2010, 2014; Wells et al. 2009, to name just a few) have shown groups of participants to differ significantly in ease and speed of processing and in the use of a wide range of constructions that vary in size, schematicity, complexity, and dispersion. Importantly, these differences appear to be related to differences in people's experiences with language.

Now, given that no two speakers are identical in their language use and language exposure, also *within* groups of participants variation is to be expected. Street & Dąbrowska (2010, 2014), in their studies on education-related differences in comprehension of the English passive construction, note that there are considerable differences in performance within the group of less educated participants, but they do not examine this in more detail. An interesting study that does zoom in on individual speakers is Barlow's (2013) investigation of the speech of six White House Press Secretaries answering questions at press conferences. While the content changes across the different samples and different speakers, the format is the same. Barlow analyzed bigrams and trigrams (e.g. *well I think, if you like*) and part-of-speech bigrams (e.g. first person plural

[10] The importance of accumulated linguistic experiences in the construction of cognitive representations is acknowledged in various fields of research, for example in work on the categorization of sounds (e.g. Goudbeek et al. 2009; Kuhl 2000).

personal pronoun + verb). He found individual differences, not just in the use of a few idiosyncratic phrases but in a wide range of core grammatical constructions.

As Barlow (2013) used multiple speech samples from each press secretary, taken over the course of several months, he was able to examine variation between and within speakers. He observed that the inter-speaker variability was greater than the intra-speaker variability, and the frequency of use of expressions by individual speakers diverged from the average. Barlow thus exemplifies one way of investigating the third type of variation: variation across time.

If you collect data from a language user on a particular linguistic item at different points in time, you may observe variation from one moment to the other. The degree of variation will depend on the type of item that is investigated and on the length of the interval. For various types of items there are clear indications of change throughout one's life, as language acquisition, attrition, and training studies show (e.g. Baayen et al. 2017; De Bot & Schrauf 2009; N. Ellis 2002). While this may seem self-evident with respect to neologisms, and words and phrases that are part of a register one becomes familiar with or ceases to use, change has also been observed for other aspects of language. Eckert (1997) and Sankoff (2006), for instance, describe how speakers' patterns of phonetic variation can continue to change throughout their lifetime.

Also in a much shorter time frame, the use of a linguistic item by a single speaker may vary. Case studies involving relatively spontaneous speech, as well as large-scale investigations involving elicited speech, demonstrate an array of structured variation available to an individual speaker. This variation is often related to stylistic aspects, audience design, and discourse function. Labov (2001: 438-445) describes how the study of the speech of one individual in a range of situations shows clear differences in the vowels' formant values depending on the setting. Sharma (2011) compares two sets of data from a young British-born Asian woman in Southall: data from a sociolinguistic interview and self-recorded interactional data covering a variety of communicative settings. Sharma reports how the latter, but not the former, revealed strategically 'compartmentalized' variation. The informant was found to use a flexible and highly differentiated repertoire of phonetic and lexical variants in managing multiple community memberships. The variation observed may follow from deliberate choices, as well as automatic alignment mechanisms (Garrod & Pickering 2004).

Variation within a short period of time need not always involve differences in style and setting. Sebregts (2015) reports on individual speakers varying between different realizations of /r/ within the same communicative setting and the same linguistic context. He conducted a large-scale investigation into the sociophonetic, geographical, and linguistic variation found with Dutch /r/.[11] In 10 cities in the Netherlands and Flanders, he asked approximately 40 speakers per city to perform a picture naming task and to read aloud a word list. The tasks involved 43 words that represent different phonological contexts in which /r/ occurs. Sebregts observed interesting patterns of variation between and within participants. In each of the geographical communities, there were differences between the individual speakers, some of them realizing /r/ in a way that is characteristic of another community. Furthermore, speaker-internal variation was found to be high. In part, this variation was related to the phonological environment in which /r/ appeared. In addition, participants seemed to have different variants at their disposal for the realization of /r/ in what were essentially the same contexts. Some Flemish speakers, for example, alternated between alveolar and uvular *r* within the same linguistic context, in the course of a five-minute elicitation task.

As Sebregts made use of two types of tasks –picture naming and word list reading– he examined not just variation across items, participants, and time, but also possible variation across methods. In his study, there were no significant differences in speakers' performance between the two tasks. His tasks thus yielded converging evidence: the results obtained via one method were confirmed by those collected in a different way. This increases the reliability of the findings. If there were to be differences, these are at least as important and interesting. Different types of data may display meaningful differences as they tap into different aspects of language use and linguistic knowledge. Methods can thus complement each other and offer a fuller picture (e.g. Chaudron 1983; Flynn 1986; Nordquist 2009; Schönefeld 2011; Kertész et al. 2012).

A growing number of studies combine various kinds of data (see Arppe et al. 2010; Gilquin & Gries 2009; Hashemi & Babaii 2013 for examples and critical discussions of the current practices). Some investigations make use of fundamentally different types of data. For instance, quantitative data can be complemented with qualitative data, to gain an in-depth understanding of particular behavior. An often-used combination is that of corpus-based and experimental evidence, to investigate how frequency patterns in spontaneous speech correlate with processing speed or metalinguistic judgments (e.g. Mos et al. 2012). Alternatively, two versions of the same experimental task can be administered, to assess possible effects of the design. For example, participants may be asked to express judgments on different kinds of ratings scales (e.g. a

[11] Note that the /r/ sound may be more naturally variable than many other sounds. As Sebregts (2015: 1) remarks: "The realisation of /r/ in Dutch is a particularly striking example of multidimensional variability".

binary scale, a Likert scale, and an open-ended scale constructed in Magnitude Estimation), to see whether the scales differ in perceived ease of use and expressivity, and in the judgment data they provide (e.g. Bader & Häussler 2010; Langsford et al. 2018; Preston & Colman 2000).

In sum, there are various indications that there is meaningful variation in the production and perception of language, and that this variation can inform theories on language processing and linguistic representations. We will demonstrate how to measure the different types of variation, and how to determine which variation can be considered informative. We do this by investigating metalinguistic judgments in combination with corpus frequency data. Judgment tasks form an often-used method in linguistics. They enable researchers to gather data on phenomena that are absent or infrequent in corpora. Furthermore, in comparison to psycholinguistic processing data, untimed judgments have the advantage of hardly being affected by factors like sneezing, a lapse of attention, or unintended distractions, as participants have ample time to reflect on the stimuli. This is not to say that untimed judgments are not subject to uncontrolled or uncontrollable factors at all (see for instance Birdsong 1989: 62-68), but they can form a valuable complement to time-pressured performance data (e.g. R. Ellis 2005). Another advantage is that it is relatively easy and cheap to conduct a judgment task with large numbers of participants. It is therefore not surprising that countless researchers make use of judgment data in the investigation of phenomena ranging from syntactic patterns (e.g. Keller & Alexopoulou 2001; Meng & Bader 2000; Sorace 2000; Schütze 1996; Sprouse & Almeida 2012; Theakston 2004) to formulaic language (e.g. N. Ellis & Simpson-Vlach 2009), collocations and constructions (Granger 1998; Gries & Wulff 2009). Nonetheless, not much is known about the degrees of variation in judgments – especially the variation across participants and across time, and the extent to which this is influenced by the design of the task. Typically, participants complete a judgment task just once, and the reports are confined to mean ratings, averaging over participants. Some studies (e.g. Langsford et al. 2018) do examine test-retest reliability of judgments expressed on various scales, thus examining variation across time and across methods, but all analyses are performed on mean ratings. We will demonstrate how all four types of variation can be investigated in judgment data, and how they can be used as sources of information.

3.2 Outline of the present research

To investigate variation in judgments across items, participants, time, and methods, we had native speakers of Dutch rate the familiarity of prepositional phrases such as *in de tuin* ('in the garden') and *rond de ingang* ('around the entrance') twice within the space of one to two weeks, using either Magnitude

Estimation or a 7-point Likert scale. While all phrases could potentially be used in everyday life, they differ in the frequency with which they occur in Dutch corpora, covering a large range of frequencies (see Section 3.3.3). The frequency of occurrence of such word sequences has been shown to affect the speed with which they are recognized and produced (e.g. Arnon & Snider 2010; Tremblay & Tucker 2011; Chapter 4), and we expect usage frequency to be reflected in familiarity ratings (cf. Balota et al. 2001; Popiel & McRae 1988; Shaoul et al. 2013). Given the gradual differences in frequency of occurrence between items, the familiarity judgments are likely to exhibit gradience as well. As we are interested in individual differences, we opted for two rating scales that allow individual participants to express such gradience (see Langsford et al. 2018 for a comparison of Likert and Magnitude Estimation scales with forced choice tasks that require averaging over participants; see Colman et al. 1997 for a comparison of data from 5- and 7-point rating scales).

By contrasting the degree of variation across participants with the degree of variation within participants, we can gain insight into the extent to which variation across speakers is meaningful. Participants perform the same judgment task twice within a time span short enough for the construct that is being tested not to have changed much, yet long enough for the respondents not to be able to recall the exact scores they assigned the first time. If each individual's judgment is fairly stable, while there is consistent variation across participants, then this shows that there are stable differences between participants in judgment. If individuals' judgments are found to vary from one moment to the other, this gives rise to another important question: Does this mean that judgments are fundamentally noisy, or is the variability a genuine characteristic of people's cognitive representations, requiring to be investigated and accounted for?

In disciplines other than linguistics, there is plenty of research taking rating scale measurements several days, weeks, or months apart (see, for instance, Ashton 2000; Churchill & Peter 1984; Jiang & Cillessen 2005; Paiva et al. 2014; VanGeest et al. 2002). Also in linguistics there are a number of studies in which participants performed (part of) a judgment task twice, some of which show judgments to be unstable (e.g. Birdsong 1989; R. Ellis 1991; Johnson et al. 1996; Tabatabaei & Dehghani 2012). Most of this research has been conducted with second language learners. Important to note is that these studies offered few response options (either binary, or acceptable/unacceptable/unsure), and the stimuli consisted of sentences. This likely influences the stability of the judgments. A binary response scale may not fit well with people's perceptions of acceptability. As Birdsong (1989: 166) puts it: "Not all grammatical sentences are perceived as equally 'good', and not all ungrammatical sentences are perceived as equally 'bad'" (also see Wasow & Arnold 2005). If you consider a stimulus to be of medium acceptability, it is not surprising that you will classify it as acceptable on one occasion and as unacceptable on another. It has been argued that more than three response options are needed to achieve stable participant responses (Preston & Colman 2000; Weng 2004). Furthermore, in the majority of the test-retest studies participants were asked to judge sentences. If language users do not store representations of entire sentences, it may be harder to assess them in the exact same way on different occasions. Consequently, these studies do not answer the question how much variation is to be expected when adult native speakers perform the same metalinguistic judgment task twice within a couple of weeks, rating phrases that may be used in everyday life on a scale that allows for more fine-grained distinctions.

The set-up of our study enabled us to compare the variation across participants with the variation across time, and to relate each of these to corpus-based frequencies of the phrases. In addition, we examined variation across methods. To be precise, we measured the four types of variation discussed in Section 3.1 and used those to test four hypotheses regarding linguistic representations and metalinguistic knowledge and to answer an as yet open question with respect to the variation across rating methods.

Hypothesis I Variation across items correlates with corpus frequencies

Rated familiarity indexes the extent and type of previous experience someone has had with a given stimulus (Gernsbacher 1984; Juhasz et al. 2015). If you are to judge the familiarity of a word string, your assessment is taken to rest on frequency and similarity to other words, constructions, or phrases (Bybee 2010: 214). Therefore, participants' ratings are expected to correlate with corpus frequencies – not perfectly, though, since a corpus is not a perfect representation of an individual participant's linguistic experiences. So, the first hypothesis will be borne out if variation across items is found that can be predicted largely from the independent variable: corpus frequencies.

Hypothesis II Variation across participants is smaller for high-frequency phrases than for low-frequency phrases

The more frequent the phrase, the more likely that it is known to many people. The use of words tends to be 'bursty': when a word has occurred in a text, you are more likely to see it again in that text than if it had not occurred (Altmann et al. 2011; Church & Gale 1995). The occurrences of stimuli with low corpus frequencies are likely to be clustered in a small number of texts. As such, they may be fairly common for some people, while others virtually never use it. Consequently, familiarity ratings for these phrases will differ more across participants.

Hypothesis III Variation across time is smaller for high-frequency phrases than for low-frequency phrases

In judging familiarity, a participant will activate potential uses of a given stimulus. The number and kinds of usage contexts and the ease with which they come to mind influence familiarity judgments. The item's frequency may affect the ease with which exemplars are generated. For low-frequency phrases, the number and type of associations and exemplars that become activated are likely to differ more from one moment to the other, resulting in variation in judgments across time.

Hypothesis IV The variation across participants is larger than the variation across time

For this study's set of items and test-retest interval, the variation in judgment across participants is expected to be larger than the variation within one person's ratings across time. As the phrases may be used in everyday life, the raters had at least 18 years of linguistic experiences that have contributed to their familiarity with these word strings. From that viewpoint, two weeks is a relatively short time span, and there is no reason to assume that the use of the word combinations under investigation, or participants' mental representations of these linguistic units, changed much in two weeks.

Question To what extent is there variation across rating methods?

As for possible variation across rating methods, different hypotheses can be formulated. Magnitude Estimation (ME) differs from Likert scales in that it offers distinctions in ratings that are as fine-grained as participants' capacities allow (Bard et al. 1996). Participants create their own scale of judgment, rather than being forced to use a scale with a predetermined, limited number of values of which the (psychological) distances are unknown. According to some researchers (e.g. Weskott & Fanselow 2011), Magnitude Estimation is more likely to produce large variance than Likert scale or binary judgment tasks, due to the increased number of response options. However, several other studies (e.g. Bader & Häussler 2010; Bard et al. 1996; Wulff 2009) provide evidence that Magnitude Estimation yields reliable data, not different from those of other judgments tasks, and that inter-participant consistency is extremely high.

One could even argue that judgments expressed by means of Magnitude Estimation will display less variation across time than Likert scale ratings. As ME allows participants to distinguish as many degrees of familiarity as they feel relevant, there is likely to be a better match between perceived familiarity and the ratings one assigns (cf. Preston & Colman 2000). A participant may have the feeling that the level of familiarity of an item corresponds to 4.5 on a 7-point scale, but this is not a valid response option on this scale. It is very well possible that this participant then rates the item as 4 on one occasion and as 5 on another occasion. If participants are free to choose the number of degrees that are distinguished, they can assign the rating 4.5 on both occasions. Moreover, the self-construal of a rating scale may involve more conscious processing and evaluation of the stimulus items. This could lead to stronger memory traces and therefore a higher correspondence in ratings across time.

3.3 Method

3.3.1 Design

In order to examine degrees of variation in familiarity judgments for prepositional phrases with a range in frequency, and the influence of using a Likert vs a Magnitude Estimation scale, a 2 (Time) x 2 (RATINGSCALE) design was used. 91 participants rated 79 items twice within the space of one to two weeks. As can be observed from Table 3.1, half of the participants gave ratings on a 7-point Likert scale at Time 1; the other half used Magnitude Estimation. At Time 2, half of the participants used the same scale as at Time 1, and the other half was given a different scale. This allowed us to investigate variation across items, across participants, across time, and across methods.

Table 3.1 The number of participants that took part in the four experimental conditions.

Rating scale at Time 1	Rating scale at Time 2	Participants N
Likert	Likert	24
Likert	Magnitude Estimation	22
Magnitude Estimation	Likert	22
Magnitude Estimation	Magnitude Estimation	23

3.3.2 Participants

The group of participants consisted of 91 persons (63 female, 28 male), mean age 27.1 years (*SD* = 11.9, age range: 18 - 70). The four conditions did not differ in terms of participants' age (*F*(3, 87) = 0.20, *p* = .89) or gender (χ^2(3) = 1.83, *p* = .63). All participants were native speakers of Dutch. A large majority (viz. 82 participants) had a tertiary education degree; 9 participants had had intermediate vocation education. Educational background did not differ across conditions (χ^2(6) = 3.57, *p* = .73).

3.3.3 Stimulus items

Participants were asked to rate 79 Prepositional Phrases (PPs) consisting of a preposition and a noun, and in a majority of the cases an article (i.e. 52 phrases with the definite article *de*; 16 with the definite article *het*; 11 without an article). The items cover a wide range of frequency (from 1 to 14688) in a subset of the corpus SoNaR consisting of approximately 195.6 million words.[12] The phrases and the frequency data can be found in Appendices 3.1 and 3.2.

The word strings were presented in isolation. Since all stimuli constitute phrases by themselves, they form a meaningful unit even without additional context. In a previous study into the stability of Magnitude Estimation ratings of familiarity (Chapter 2), we investigated possible effects of context by presenting prepositional phrases both in isolation and embedded in a sentence. The factor CONTEXT did not have a significant effect on familiarity ratings, nor on the degrees of variation across and within participants.

3.3.4 Procedure

The items were presented in an online questionnaire form (using the Qualtrics software program) and this was also the environment within which the ratings were given. The experiment was conducted via the internet.[13] Participants received a link to a website. There they were given more information about the study and they were asked for consent. Subsequently, they were asked to provide some information regarding demographic variables (age, gender, language background, educational background). After that, it was explained that their task was to indicate how familiar various word combinations are to them. In line with earlier studies using familiarity ratings (Juhasz et al. 2015; Williams & Morris 2004), our instructions read that the more you use and encounter a particular word combination, the more familiar it is to you, and the higher the score you assign to it.

In the Likert scale condition, participants were presented with a prepositional phrase together with the statement ‘This combination sounds familiar to me’ (*Deze combinatie klinkt voor mij vertrouwd*) and a 7-point scale, the endpoints of which were marked by the words ‘Disagree’ and ‘Agree’ (*Oneens* and *Eens*). Participants were shown one example. After that, the experiment started.

[12] SoNaR is a balanced reference corpus of contemporary written standard Dutch (Oostdijk et al. 2013). The subset we used consists of texts originating from the Netherlands (143.8 million words) and texts originating either from the Netherlands or Belgium (51.8 million words).

[13] Balota et al. (2001) found that familiarity ratings from a web-based task were strongly correlated with ratings from laboratory tasks.

When participants were to use Magnitude Estimation, they were first introduced to the notion of relative ratings through the example of comparing the size of depicted clouds and expressing this relationship in numbers. In a brief practice session, participants gave familiarity ratings to word combinations that did not comprise prepositional phrases (e.g. *de muziek klinkt luid* 'the music sounds loud'). Before starting the main experiment, they were given advice not to restrict their ratings to the scale used in the Dutch grading system (1 to 10, with 10 being a perfect score), not to assign negative numbers, and not starting very low, to allow for subsequent lower ratings. At the start of the experiment, participants rated the phrase *tegen de avond* ('towards the evening'). This phrase was taken from the middle region of the frequency range, as this may stimulate sensitivity to differences between items with moderate familiarity (Sprouse 2011). Then, they compared each successive stimulus to the reference phrase ('How do you rate this combination in terms of familiarity when comparing it with the reference combination?' *Hoe scoort deze combinatie op vertrouwdheid wanneer je deze vergelijkt met de referentiecombinatie?*).

The stimuli were randomized once. The presentation order was the same for all participants, in both sessions, to ensure that any differences in judgment are not caused by differences in stimulus order (cf. Sprouse 2011). Midway, participants were informed that they had completed half of the task and they were offered the opportunity to fill in remarks and questions, just like they were at the end of the task.

All participants completed the experiment twice, with a period of one to two weeks between the first and second session. They knew in advance that the investigation involved two test sessions, but not that they would be doing the same task twice. The time interval ranged from 4 to 15 days (M = 7, SD = 3.11). The four experimental conditions did not differ in terms of time interval (F(3, 87) = 0.28, p = .84). After four days, people are not expected to be able to recall the exact scores they assigned to each of the 79 stimuli.

3.3.5 Data transformations

For each participant, the ratings provided within one session were converted into Z-scores to make comparisons of judgments and variation possible. By converting into Z-scores, a score of 0 indicates that a particular item is judged by a participant to be of average familiarity compared to the other items. For each item, Appendix 3.2 lists the mean of the Z-scores of all participants for that item, and the standard deviation. The Z-score transformation is common in judgment studies (Bader & Häussler 2010; Schütze & Sprouse 2013), as it involves no loss of information on ranking, nor at the interval level. It does entail the loss of information about absolute familiarity and developments in absolute familiarity over time that is present in the data from the Likert scale condition. However, absolute familiarity is of secondary importance in this study. A direct comparison of the different response variables, on the other hand, is at the heart of the matter, and the use of Z-scores enables us to make such a comparison. To assess the consequences of using Z-scores, we also performed all analyses using raw instead of standardized Likert scores, applying mixed ordinal regression to the Likert scale data, and linear mixed-effects models to the ME data. This did not yield substantially different findings. We will come back to differences between Likert and ME ratings, and advantages and disadvantages of each of those, in the discussion (Section 3.5).

To investigate variation across time, a participant's Z-score for an item in the second session was deducted from the score in the first session. The difference (i.e. Δ-score) provides insight in the extent to which a participant rated an item differently over time (e.g. if a participant's rating for *naar huis* yielded a Z-score of 1.0 in the first session, and 0.5 in the second, the Δ-score is 0.5; if it was 1.0 the first time, and 1.5 the second time, the Δ-score is also 0.5, as the variation across time is of the same magnitude). Given that participants who used Magnitude Estimation constructed a scale at Time 1 and a new one at Time 2, ratings had to be converted into Z-scores at Time 1 and Time 2 separately. Consequently, we cannot determine whether participants might have considered *all* stimuli more familiar the second time (something which will be addressed in Section 3.5).

In order to relate variation in judgments to frequency of the phrases, frequency counts of the exact word string in the SoNaR-subset were queried and the frequency of occurrence per million words in the corpus was logarithmically transformed to base 10. The same was done for the frequency of the noun (lemma search).[14] To give an example, the phrase *naar huis* occurred 14,688 times, which corresponds to a log-transformed frequency score of 1.88. The lemma frequency of the noun, which encompasses occurrences of *huizen, huisje,*

[14] Knowledge about the patterns of co-occurrence of linguistic elements is part of our mental representations of language. Such knowledge is taken to inform familiarity judgments. It also enables us to generate expectations, which in turn affects the effort it takes to process the subsequent input (Huettig 2015). Word predictability is commonly expressed by means of the metrics entropy (which expresses the uncertainty at position t about what will follow) and surprisal (which expresses how unexpected the actually perceived word w_{t+1} is), estimated by language models trained on text corpora (Levy 2008). Entropy and surprisal have been used successfully in models that predict speed and ease of processing (e.g. Baayen et al. 2011; Linzen & Jaeger 2016). These metrics are not taken into account in the present study, as we do not examine processing costs. We do so in another paper, in which we examine individual differences in experiences, expectations, and processing speed (Chapter 4).

huisjes in addition to *huis*, amounts to 84,918 instances. This corresponds to a log-transformed frequency score of 2.64. Figure 3.1 shows the positions of the stimuli on the phrase frequency scale and the lemma frequency scale; Appendix 3.2 lists for all stimuli the raw and the log-transformed frequencies. As can be observed from Figure 3.1, for low-frequency PPs, the frequency of the noun varies considerably (compare, for example, items 10 and 12). High noun frequency (like in item 12) here indicates that the noun also occurs in phrases other than the one we selected as a stimulus. Such phrases may come to mind when rating the stimulus. If some of them are considered more familiar, the score assigned to the stimulus is likely to be lowered. The high-frequency phrases in our stimulus set have fewer 'salient competitors'. They tend to be the most common phrase comprising the given noun. Consider as an example the noun *bad* ('bath', LOGFREQN 1.52). When used together with a preposition, the phrase *in bad* (item 54) is the most frequent combination (logFreqPP 0.81). Other phrases are much less frequent: *uit bad* (logPP -0.38), *met bad* (logPP -1.18).

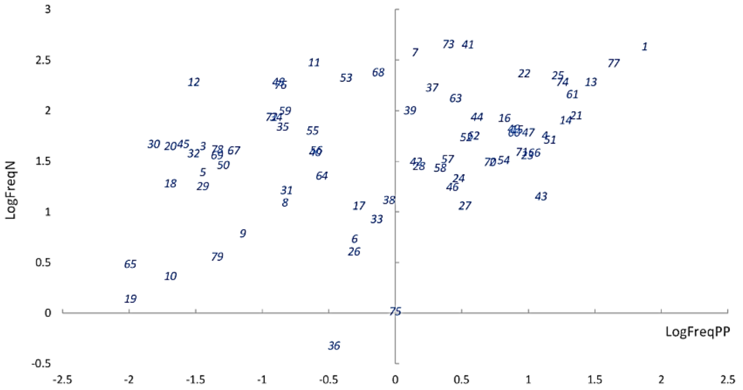


Figure 3.1 Scatterplot of the relationship between the log-transformed corpus frequency per million words of the PP and that of the N (*r* = .39). The numbers 1 to 79 identify the individual stimuli (see Appendices).

3.3.6 Statistical analyses

Using linear mixed-effects models (Baayen et al. 2008), we investigated to what extent the familiarity judgments can be predicted by corpus frequencies, and whether this differs per session and/or per rating scale. Mixed-models obviate the necessity of prior averaging over participants and/or items, enabling the researcher to model the individual response of a given participant to a given item (Baayen et al. 2008). Appendix 3.3 describes our implementation of this statistical technique (i.e. fixed effects, random effects structures, estimation of confidence intervals). If the resulting model shows that frequency has a significant effect, this is in line with our first hypothesis, which states that there is variation across items in familiarity ratings that can be predicted largely from corpus frequencies.

We used standard deviation as a measure of variation across participants. Plotting the standard deviations against the stimuli's corpus frequencies, we examined whether there is a relationship between phrase frequency and the variation in judgment across participants. We hypothesized that high-frequency phrases display less variation across participants than low-frequency phrases.

Variation across time was investigated in two ways. First, we inspected the extent to which the judgments at Time 2 correlate with the judgments at Time 1, by calculating the correlation between a participant's Z-scores across sessions. The Z-scores preserve information on ranking and on the intervals between the raw scores. High correlation scores thus indicate that there is little variation across time in these respects. Subsequently, we ran linear mixed-effects models on the Δ-scores, to determine which factors influence variation across time. As described in Section 3.3.5, the Δ-scores quantify the extent to which a participant's rating for a particular item at Time 2 differs from the rating at Time 1. The details of the modeling procedure are also described in Appendix 3.3. In order for our third hypothesis to be confirmed, phrase frequency should prove to have a significant negative effect, such that higher phrase frequency entails less variation in judgment across time.

Then we compared the variation within participants across time with the variation across participants. The latter was hypothesized to be larger than the former. If that is the case, participants' ratings at Time 1 should be more similar to their own ratings at Time 2 than to the other participants' ratings at Time 2. To test this, we compared each participant's self-correlation to the correlation between that person's ratings at T1 and the group mean at T2, by means of the procedure described by Field (2013: 287).[15] If the latter is significantly higher than the former, the fourth hypothesis is confirmed.

[15] Field (2013: 287) describes how one can test by means of a *t*-statistic (Chen & Popovich 2002) whether a difference between two dependent correlations from the same sample is significant. To test whether the relationship between a participant's scores at Time 2 (x) and that participant's scores at Time 1 (y) is stronger than the relationship between the group mean at Time 2 (z) and that participant's scores at Time 1 (y), the *t*-statistics is computed as:
$t_{\text{Difference}} = (r_{xy} - r_{zy}) * \surd\ (((n - 3)(1+ r_{xz})) \ / \ (2(1 - r^2_{xy} - r^2_{xz} - r^2_{zy} + 2*r_{xy}*r_{xz}*r_{zy})))$
The resulting value is checked against the appropriate critical values. For a two-tailed test with 76 degrees of freedom, the critical values are 1.99 ($p < .05$) and 2.64 ($p < .01$).

In order to ascertain to what extent there is variation across rating methods, we examined the role of the factor RatingScale in the linear mixed-effects models, and the extent to which the patterns in the standard deviations as well as the Time1–Time2 correlations vary depending on the rating scale that is used. To conclude that the scales yield different outcomes, the standard deviations and correlation scores should be found to differ across methods, and/or the factor RatingScale should prove to have a significant effect, or enter into an interaction with another factor, in the mixed-models.

3.4 Results

3.4.1 Relating familiarity judgments to corpus frequencies and rating scale

Participants discerned various degrees of familiarity. In the Likert scale conditions, participants could distinguish maximally seven degrees. On average, they discerned 6.4 degrees of familiarity (Likert Time 1: *M* = 6.3, *SD* = 1.2, range: 2-7; Likert Time 2: *M* = 6.5, *SD* = 1.0, range: 2-7). In the Magnitude Estimation conditions, participants could determine the number of response options themselves. On average, they discerned 12.0 degrees of familiarity (ME Time 1: *M* = 12.6, *SD* = 6.3, range: 3-35; ME Time 2: *M* = 11.4, *SD* = 4.4, range: 3-22).

From a usage-based perspective, perceived degree of familiarity is determined to a large extent by usage frequency, which can be gauged by corpus frequencies. By means of linear mixed-effects models, we investigated to what extent the familiarity judgments can be predicted by the frequency of the specific phrase (LogFreqPP) and the lemma-frequency of the noun (LogFreqN), and to what degree the factors RatingScale (i.e. Likert or Magnitude Estimation), Time (i.e. first or second session), and the order in which the items were presented exert influence. We incrementally added predictors and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in familiarity judgments. A detailed description of this model selection procedure can be found in Appendix 3.3. The interaction term LogFreqPP x LogFreqN did not contribute to the fit of the model. Furthermore, none of the interactions of Time and the other variables was found to improve goodness-of-fit. As for PresentationOrder, only the interaction with RatingScale contributed to explaining variance. The resulting model is summarized in Table 3.2. The variance explained by this model is 57% (R^2m = .36, R^2c = .57).[16]

[16] R2m (marginal R^2 coefficient) represents the amount of variance explained by the fixed effects; R2c (conditional R^2 coefficient) is interpreted as variance explained by both fixed and random effects (i.e. the full model) (Johnson 2014).

Table 3.2 Estimated coefficients, standard errors, and 95% confidence intervals for the mixed-model fitted to the standardized familiarity ratings.

	B	*SE b*	*t*	*95 % CI*
Intercept	0.00	0.05	0.00	*-0.10, 0.09*
LogFreqPP	**0.59**	**0.05**	**10.85**	***0.47, 0.69***
LogFreqN	-0.01	0.05	-0.10	*-0.11, 0.10*
RatingScale	-0.00	0.02	-0.01	*-0.04, 0.03*
RatingScale x LogFreqPP	0.01	0.02	0.50	*-0.03, 0.05*
RatingScale x LogFreqN	0.04	0.02	1.68	*-0.01, 0.08*
PresentationOrder	-0.04	0.05	-0.80	*-0.14, 0,05*
PresentationOrder x RatingScale	-0.03	0.02	-1.46	*-0.06, 0.01*

Note: Significant effects are printed in bold.

The factor RatingScale did not have a significant effect, indicating that familiarity ratings expressed on a Magnitude Estimation scale do not differ systematically from familiarity ratings expressed on a Likert scale. Furthermore, the factor RatingScale did not enter into any interactions with other factors. This means that the role of these factors does not differ depending on the scale used.

As can be observed from Table 3.2, just one factor proved to have a significant effect: LogFreqPP. Only the frequency of the phrase in the corpus significantly predicted judgments, with higher frequency leading to higher familiarity ratings, as can be observed from Figure 3.2. This phrase frequency effect was found both in Likert and ME ratings, at Time 1 as well Time 2.

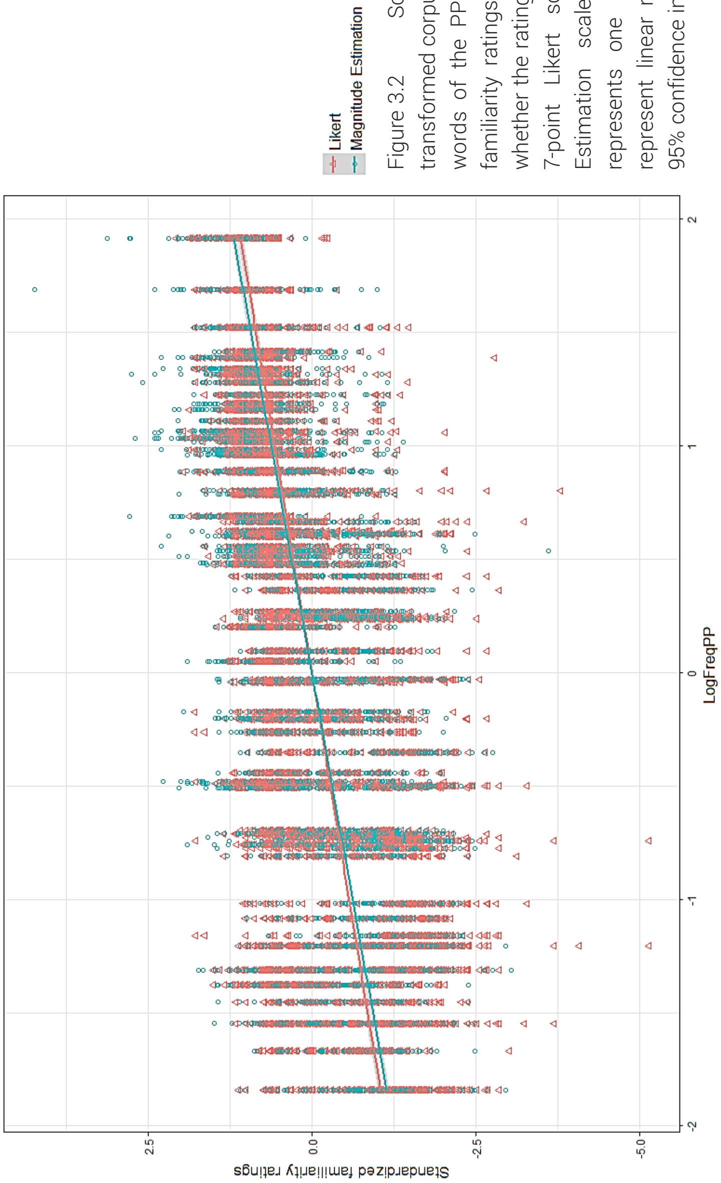


Figure 3.2 Scatterplot of the log-transformed corpus frequency per million words of the PP and the standardized familiarity ratings, split up according to whether the ratings were expressed on a 7-point Likert scale or a Magnitude Estimation scale. Each circle/triangle represents one observation; the lines represent linear regression lines with a 95% confidence interval.

3.4.2 Variation across participants

Given that people differ in their linguistic experiences, familiarity with particular word strings was expected to vary across participants, and the differences were hypothesized to be larger in phrases with low corpus frequencies compared to high-frequency phrases. The standard deviations listed in Appendix 3.2 quantify per item the amount of variation in judgment across participants. Figure 3.3 plots these standard deviations against the corpus frequencies of the phrases. Low-frequency phrases tend to display more variation in judgment across participants than high-frequency phrases, as evidenced by higher standard deviations. This holds for Likert ratings more so than for ME ratings.

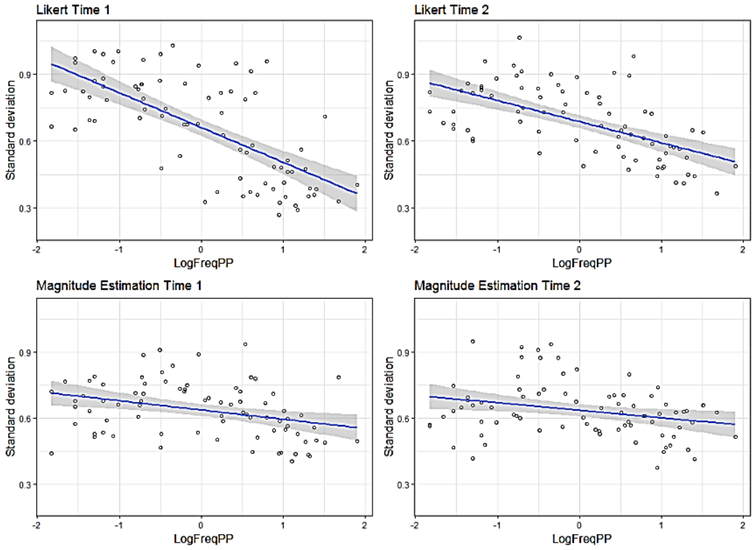


Figure 3.3 Scatterplots of the standard deviations in relation to the log-transformed corpus frequency per million words of the PP. The lines represent linear regression lines with a 95% confidence interval around it.

3.4.3 Variation across time

To examine variation across time, we calculated the correlation between the ratings assigned at Time 1 and those assigned at Time 2. When averaging over participants, the ratings are highly stable, regardless of the scales that were used. Per condition, we computed mean ratings for each of the 79 items at Time 1, and likewise at Time 2. The correlation between these two sets of mean ratings is nearly perfect in all four conditions (see Table 3.3).

Table 3.3 Correlation of mean standardized ratings at Time 1 and Time 2 (Pearson's *r*).

Time 1	Time 2	Correlation mean ratings T1 – T2	*95 % CI*
Likert	Likert	.97	*.96, .98*
Likert	ME	.96	*.94, .97*
ME	Likert	.98	*.97, .98*
ME	ME	.98	*.97, .99*

We also examined the stability of individual participants' ratings. For each participant we computed the correlation between that person's judgments at Time 1 and that person's judgments at Time 2. This yielded 91 correlation scores that range from -.31 to .90, with a mean correlation of .70 (*SD* = .20). The four conditions do not differ significantly in terms of intra-individual stability (*H*(3) = 4.76, *p* = .19). If anything, the ME-ME condition yields slightly more stable judgments than the other conditions, as can be observed from Table 3.4 and Figure 3.4.

Table 3.4 Distribution of individual participants' Time 1 – Time 2 correlation (Pearson's *r*) of standardized scores.

Time 1	Time 2	Average of individual participants' correlation (*SD*)	Range
Likert	Likert	.67 (.27)	-.31 – .87
Likert	ME	.66 (.21)	-.01 – .86
ME	Likert	.72 (.14)	.38 – .87
ME	ME	.76 (.11)	.45 – .90

There are three participants whose ratings at Time 2 do not correlate at all with their ratings on the same items, with the same instructions and under the same circumstances a few weeks earlier (*r* < .20). Two of them were part of the Likert-Likert group; one of them belonged to the Likert-ME group.[17] The majority of the participants had much higher scores, though, and this holds for all conditions. In total, 7.7% of the participants (N = 7) had self-correlation scores ranging from .20 to .50; 34.1% (N = 31) had scores ranging from .51 to .75; 54.9% (N = 50) had

[17] Low self-correlation scores are not related to educational background. The three participants with self-correlation scores below .20 had intermediate vocational education, higher vocational education, and higher education. As regards the group with self-correlation scores ranging from .20 to .49, one participant had intermediate vocational education, and the others had a tertiary education degree.

scores ranging from .76 to .90. Still, none of the participants is as stable in their ratings as the aggregated ratings presented in Table 3.3.

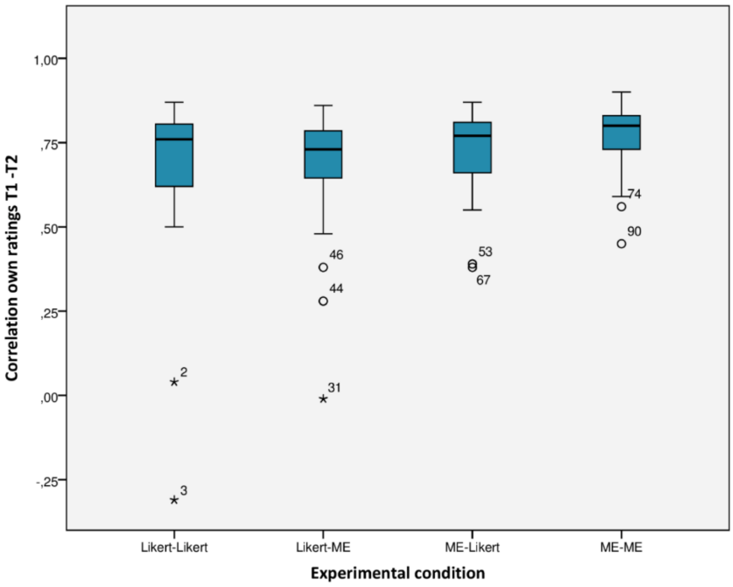


Figure 3.4 Boxplot of participants' correlation of their own standardized ratings (Pearson's *r*, Time 1 – Time 2).

3.4.4 Variation across time vs. variation across participants

If participants' ratings at Time 1 are more similar to their own ratings at Time 2 than to the other participants' ratings at Time 2, this indicates that the variation across participants is larger than variation across time. We compared each participant's self-correlation to the correlation between that person's ratings at T1 and the group mean at T2 (following Field 2013: 287). For 8 participants, self-correlation was significantly higher than correlation with the group mean; for 19 participants correlation with the group mean was significantly higher than self-correlation; for 64 participants there was no significant difference between the two measures. All experimental conditions showed a similar pattern in this respect.

3.4.5 Variation across time in relation to corpus frequencies and rating scale

In order to determine if familiarity ratings were stable for *certain items* more so than for others, or for one rating scale more so than for the other, we analyzed the Δ-scores using linear mixed-models (see Sections 3.3.5 and 3.3.6). To be precise, we investigated to what extent variation across time is related to frequency of the phrase and the noun and to the rating scales used at Time 1 and Time 2.[18] The resulting model is summarized in Table 3.5.

Table 3.5 Estimated coefficients, standard errors, and 95% confidence intervals for the mixed-model fitted to the log-transformed absolute Δ-scores.

	b	*SE b*	*t*	*95 % CI*
Intercept	-1.31	0.10	-12.63	-1.51, -1.10
LogFreqPP	**-0.26**	**0.06**	**-4.34**	**-0.37, -0.14**
RatingScaleT1	0.04	0.12	0.33	-0.20, 0.28
RatingScaleT2	0.18	0.12	1.52	-0.06, 0.41
LogFreqPP x RatingScaleT1	**0.17**	**0.07**	**2.53**	**0.04, 0.31**
LogFreqPP x RatingScaleT2	0.09	0.07	1.43	-0.03, 0.22

Note: Significant effects are printed in bold.

The type of scale that was used did not have a significant effect on the variation across time. Furthermore, the interaction term RatingScaleT1 x RatingScaleT2 did not contribute to explaining variance in Δ-scores (see Appendix 3.3). One may have expected ratings to be more stable if the same type of scale was used across sessions (i.e. Likert-Likert or ME-ME, rather than Likert-ME or ME-Likert). The fact that the interaction RatingScaleT1 x RatingScaleT2 did not improve model fit shows that this was not the case.

LogFreqPP proved to have a significant effect, and there was a significant interaction of LogFreqPP with RatingScaleT1. In general, higher phrase frequency led to less variation in judgment across time. However, the relationship between phrase frequency and instability in judgment was not observed in all experimental conditions (see Figure 3.5). It holds for the ratings when at Time 1 Likert-scales were used to express familiarity (i.e. the two plots on the left in Figure 3.5).

[18] As was reported in Section 3.3.4, the phrases were presented in a fixed order, the same for all participants. We tested whether there were effects of fatigue (e.g. more instability towards the end of the experiment) by including the factor PresentationOrder in the mixed-effects models. Neither PresentationOrder, nor any of the interactions of PresentationOrder and the other predictors was found to improve model fit (see Appendix 3.3).

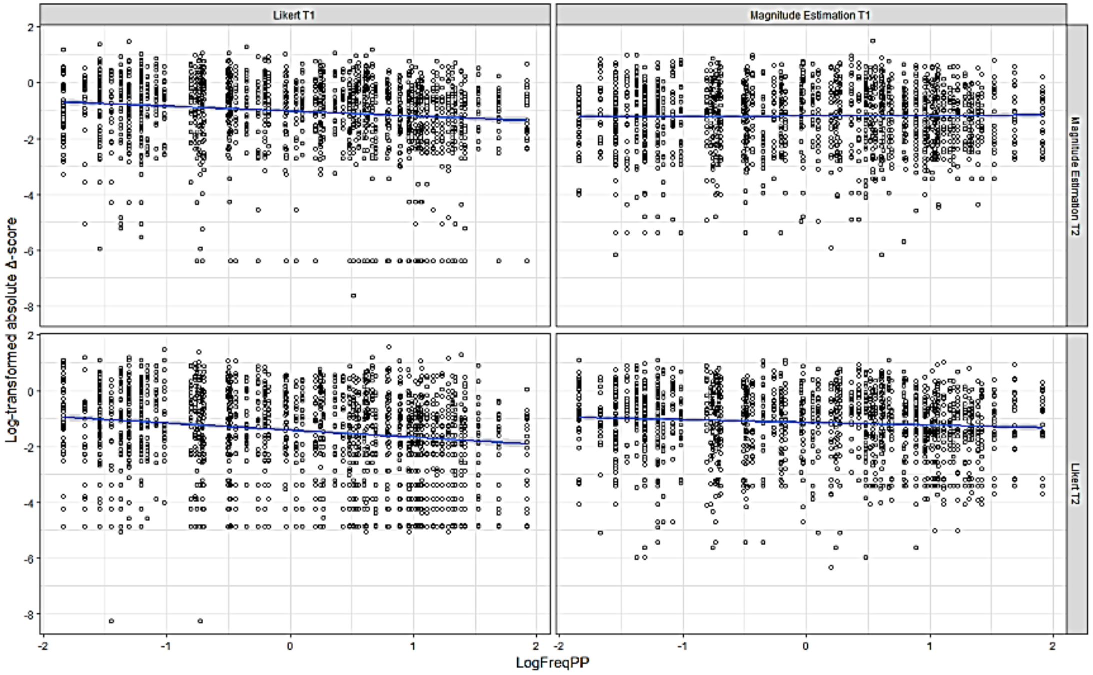


Figure 3.5 Scatterplot of the log-transformed corpus frequency per million words of the PP and the log-transformed absolute Δ-scores, per experimental condition. Each circle represents one observation; the lines represent linear regression lines with a 95% confidence interval around them. Note: The lower the log-transformed Δ-score, the more stable the judgments were. For instance, a Δ-score of 0.02 (meaning very little difference between the ratings at Time 1 and Time 2) corresponds to a log-transformed Δ-score of -3.91.

3.5 Discussion

For a long time, variation has been overlooked, ignored, looked at from a limited perspective (e.g. variation being simply the result of irrelevant performance factors), or considered troublesome in various fields of linguistics. The variation observable in metalinguistic performance made Birdsong (1989: 206-207) wonder, rather despairingly: "Should we throw up our hands in frustration in the face of individual, task-related, and situational differences, or should we blithely sweep dirty data under the rug of abstraction?" Our answer to that question is: neither of those. We argue that it is both feasible and valuable to study different types of variation. Such investigations yield a more accurate presentation of the data, and they contribute to the refinement of theories of linguistic knowledge. To illustrate this, we had native speakers of Dutch rate the familiarity of a large set of prepositional phrases twice within the space of one to two weeks, using either Magnitude Estimation or a 7-point Likert scale. This dataset enabled us to examine variation across items, variation across participants, variation across time, and variation across rating methods. We have shown how these different types of variation can be quantified and use them to test hypotheses regarding linguistic representations.

Our analyses indicate, first of all, that familiarity judgments form methodologically reliable, useful data in linguistic research. The ratings we obtained with one scale were corroborated by the ratings on the other scale (recall that there was no main effect of the factor RatingScale in the analysis of the judgments, indicating that the ratings expressed on a Magnitude Estimation scale did not differ systematically from the ratings expressed on a Likert scale). In addition, there was a near perfect Time1–Time2 correlation of the mean ratings in all experimental conditions, and the majority of the participants had high self-correlation scores. Furthermore, the data show a clear correlation between familiarity ratings and corpus frequencies. As familiarity is taken to rest on usage frequency, the ratings were hypothesized to display variation across items that could be predicted largely from corpus frequencies (but not fully, since no corpus can be a perfect representation of an individual participant's linguistic experiences, cf. Mandera et al. 2017). This prediction was borne out. Both in the Likert and in the ME condition, at Time 1 as well as at Time 2, higher phrase frequency led to higher familiarity ratings. These findings indicate that the participants performed the task properly, and that the tasks measured what they were intended to measure.

In addition to variation across items, we observed variation across participants and variation across time in familiarity ratings. These types of variation are indicative of the dynamic nature of linguistic representations. Put differently, variation is part of speakers' linguistic competence. Usage-based exemplar models naturally accommodate such variation (e.g. Goldinger 1996; Hintzman 1986; Pierrehumbert 2001). In these models, linguistic representations consist of a continually updating set of exemplars that include a large amount of detail concerning linguistic and extra-linguistic properties. An exemplar is strengthened when more and/or more recent tokens are categorized as belonging to it. Representations are thus dynamic and detailed, naturally embedding the variation that is experienced.

This variation can then be exploited by a speaker in the construction of social and geographical identities (e.g. Sebregts 2015; Sharma 2011). It can also come to the fore unintentionally, as in familiarity judgments that differ slightly across rating sessions. While the judgment task requires people to indicate the position of a given item on a scale of familiarity by means of a single value, its familiarity for a particular speaker may best be viewed as a moving target located in a region that may be narrower or wider. In that case, there is not just one true value, but a range of scores that constitute true expressions of an item's familiarity. Variation in judgment across time is not noise then, but a reflection of the dynamic character of cognitive representations as more, or less, densely populated clouds of exemplars that vary in strength depending on frequency and recency of use. While a single familiarity rating can be *a* true score, it does not offer a complete picture.[19]

This also implies that prudence is in order in the interpretation of a difference in judgment between participants on the basis of a single measurement. Such a difference cannot be taken as *the* difference in their metalinguistic representations. Not because this difference should be seen as mere noise (as Featherston 2007 contends), but because it portrays just part of the picture. It is only when you take into account the range of each individual's dynamic representations that you arrive at a more accurate conclusion. Future research should also look at mental representations of (partially) schematic constructions, including syntactic patterns, using this method. In a usage-based approach, these are assumed not to be essentially different from the lexical phrases we tested.

If you intend to measure variation across items, participants, and/or time, what kind of instrument would be most suitable? Our investigation shows that in several respects, Magnitude Estimation and a 7-point Likert scale yield similar

[19] Smits et al. (2006) proposed with respect to speech sound representations that they can be viewed as distributions. It would be interesting to investigate whether this also applies to familiarity judgments. By means of an artificial language paradigm, one would be able to control the distributional properties of the input. If metalinguistic judgments are then collected in a repeated-measures design, one can examine whether the judgments take the form of a distribution, and if so, to what extent it corresponds to the distribution in the input.

outcomes. The Magnitude Estimation ratings did not differ significantly from the ratings expressed on the Likert scale, as evidenced by the absence of an effect of the factor RatingScale in the analysis of the familiarity judgments. Both types of ratings showed a significant effect of phrase frequency. There were no significant differences between the scales in terms of Time1–Time2 correlations. Nevertheless, there are certain differences between Likert and ME ratings that deserve attention and that ought to be taken into account when selecting a particular scale.

One such difference is the possibility to determine whether participants consider the majority of items to be familiar (or unfamiliar). If most items receive a rating of 5 or more on a 7-point scale, this indicates that they are perceived as fairly familiar. ME data only show to what extent particular stimuli are rated as more familiar than others; they do not provide any information as to how familiar that is in absolute terms.

Another difference concerns the possibility to determine whether participants consider the entire set of stimuli more familiar the second time, as a result of the exposure in the test sessions. The method of Magnitude Estimation entails that the raw scores from different sessions cannot be compared directly, as a participant may construct a new scale at each occasion. Consequently, a score of 50 assigned by someone at Time 2 does not necessarily mean the same as a score of 50 assigned by that participant at Time 1: at Time 2 that participant's scale could range from 50 upwards, while 50 may have represented a relatively high score on that same person's ME scale at Time 1. Magnitude Estimation therefore requires raw scores to be converted into Z-scores for each session separately. If all items are considered more familiar at Time 2, while the range of the scores and the ranking of the items remain the same across sessions, the Z-scores at Time 1 and Time 2 will be the same. When participants use the same fixed Likert scale on both occasions, the researcher is better able to compare the raw scores directly. Although there is no guarantee that a participant interprets and uses the Likert scale in exactly the same way on both occasions, any changes are arguably limited in scope. A Likert scale thus allows you to examine whether all stimuli received a higher rating in the second session, provided that there is no ceiling effect preventing increased familiarity to be expressed for certain items. If such an analysis is of importance in your investigation, a Likert scale with a sufficient number of response options may be more useful than Magnitude Estimation. For the participants who were assigned to the Likert-Likert condition, we conducted this additional analysis, calculating Δ-scores on the basis of the raw Likert scores. This yielded 1896 Δ-scores. 48.7% of those equaled zero, meaning that a participant assigned exactly the same Likert score to a particular stimulus at Time 1 and Time 2. A further 30.6% consisted of a difference in rating across time of maximally one point on a 7-point Likert scale; 10.5% involved a difference of two points. The remaining 10.2% of the Δ-scores comprised a difference of more than two points. In 31.5% of the cases, a stimulus was rated (slightly) higher at Time 1 than at Time 2; in 19.8% of the cases, a stimulus was rated (slightly) higher at Time 2 than at Time 1.

If a researcher decides to use a Likert scale, it would be advisable to carefully consider the number of response options. When offered the opportunity to distinguish more than seven degrees of familiarity, participants in our study did so in the vast majority (83.3%) of the cases. The extent to which participants would like a scale to be fine-grained may depend on the construct that is being measured. If prior research offers little insight in this respect, researchers could conduct a pilot study using scales that vary in number of response options.

One more difference we observed between the ME scale and the Likert scale concerns the effect of phrase frequency on variation across participants and variation across time. In Likert ratings, these types of variation were more pronounced in low-frequency items than in high-frequency ones. This effect did not occur in the Magnitude Estimation ratings. While there may be explanations for the susceptibility of Likert ratings to variation among low-frequency stimuli, this is not an intentional effect of the Likert scale as a measuring instrument, and one should be aware that it might not be observed when a different type of scale is used. To fully understand this difference between Magnitude Estimation and Likert scales, more research is needed using participants whose experience with particular stimuli is known to vary. In any case, Weskott and Fanselow's (2011) suggestion that Magnitude Estimation judgments are more liable to producing variance than Likert ratings is contested by our data.

As we make a case for variation to be seen as a source of information, it remains for us to answer the question: in which cases is variation really spurious? We suggest that in untimed metalinguistic judgments variation is hardly ever noise. A typo gone unnoticed (e.g. '03' instead of '30') could be considered noise; if participants had another look, they would identify it as a mistake and correct it. In the unfortunate case that participants get bored, they might assign random scores to finish as quickly as possible. Crucially, in both cases, the ratings entered are in effect no real judgments. All variation in actual judgments stems from characteristics of language use and linguistic representations, and is therefore theoretically interesting. This is not to say that there will be no unexplained variance in the data. But instead of representing noise, this variance is information waiting to be interpreted. There are factors that have not yet been identified as relevant, as a result of which they are neither controlled for nor included in the analyses, or that we have not yet been able to operationalize. To cite Birdsong (1989: 69) once more: "Metalinguistic data are like 25-cent hot dogs: they contain meat, but a lot of other ingredients, too. Some of these ingredients resist ready identification. (…) linguistic theorists are becoming alert to the necessity of knowing what these ingredients are.” Ignoring the variation present in the data will most certainly not enhance our understanding of these ‘other ingredients’ and the way they play a part in the representation and use of linguistic knowledge. Let us explore the opportunities analyses of variance offer and realize the full potential.

Chapter 4

Abstract

While theories on predictive processing posit that predictions are based on one's prior experiences, experimental work has effectively ignored the fact that people differ from each other in their linguistic experiences and, consequently, in the predictions they generate. We examine usage-based variation by means of three groups of participants (recruiters, job-seekers, and people not (yet) looking for a job), two stimuli sets (word sequences characteristic of either job ads or news reports), and two experiments (a completion task and a Voice Onset Time task). We show that differences in experiences with a particular register result in different expectations regarding word sequences characteristic of that register, thus pointing to differences in mental representations of language. Subsequently, we investigate to what extent different operationalizations of word predictability are accurate predictors of voice onset times. A measure of a participant's own expectations proves to be a significant predictor of processing speed over and above word predictability measures based on amalgamated data. These findings point to actual individual differences and highlight the merits of going beyond amalgamated data. We thus demonstrate that is it feasible to empirically assess the variation implied in usage-based theories, and we advocate exploiting this opportunity.

This chapter is based on:
Verhagen, V., Mos, M., Backus, A. & Schilperoord, J. (2018). Predictive language processing revealing usage-based variation. *Language and Cognition, 10*(2), 329–373. https://doi.org/10.1017/langcog.2018.4

Acknowledgements
I thank Louis Onrust and Antal van den Bosch for their help in analyzing the corpus data, Sanneke Vermeulen for her help in collecting the experimental data, and Elaine Francis and two reviewers for their helpful comments on the paper we submitted to Language and Cognition.

Chapter 4 Predictive language processing revealing usage-based variation

4.1 Introduction

Prediction-based processing is such a fundamental cognitive mechanism that it has been stated that brains are essentially prediction machines (A. Clark 2013). Language processing is one of the domains in which context-sensitive prediction plays an important role. Predictions are generated through associative activation of relevant mental representations. Prediction-based processing can thus yield insight into mental representations of language. This understanding can be deepened by paying attention to variation across speakers. As yet, most investigations in this field of research suffer from a lack of attention to such variation. We will show why this is an important limitation and how it can be remedied.

A variety of studies indicate that people generate expectations about upcoming linguistic elements and that this affects the effort it takes to process the subsequent input (see Huettig 2015; Kuperberg & Jaeger 2016; Kutas, DeLong & Smith 2011 for recent overviews). One of the types of knowledge that can be used to generate expectations is knowledge about the patterns of co-occurrence of words, which is mainly based on prior experiences with these words. To date, word predictability has been expressed as surprisal based on co-occurrence frequencies in corpus data, or as cloze probability based on completion task data. Predictive language processing, then, is usually demonstrated by relating surprisal or cloze probability to an experimental measure of processing effort, such as reaction times. If a word's predictability is determined by the given context and stored probabilistic knowledge resulting from cumulative exposure, surprisal or cloze probability can be used to predict ease of processing.

Crucially, in nearly all studies to date, the datasets providing word predictability measures come from different people than the datasets indicating performance in processing tasks, and that is a serious shortcoming. Predictability will vary across language users, since people differ from each other in their linguistic experiences. The corpora that are commonly used are at best a rough approximation of the participants' individual experiences. Whenever cloze probabilities from a completion task are related to reaction time data, the experiments are conducted with different groups of participants. The studies conducted so far offer little insight into the degrees of individual variation and task-dependent differences, and they adopt a coarse-grained approach to the investigation of prediction-based processing.

The main goal of this paper is to reveal to what extent differences in experience result in different expectations and responses to experimental stimuli, thus pointing to differences in mental representations of language. This advances our understanding of the theoretical status of individual variation and its methodological implications. We use two domains of language use and three groups of speakers that can reasonably be expected to differ in experience with one of these domains. First, we examine the variation within and between groups in the predictions participants generate in a completion task. Subsequently, we investigate to what extent a participant's own expectations affect processing speed. If both the responses in a completion task and the time it takes to process subsequent input are reflections of prediction-based processing, then an individual's performance on the processing task should correlate with his or her performance on the completion task. Moreover, given individual variation in experiences and expectations, a participant's own responses in the completion task may prove to be a better predictor than surprisal estimates based on data from other people.

To investigate this, we conducted two experiments with the same participants who belonged to one of three groups: recruiters, job-seekers, and people not (yet) looking for a job. These groups can be expected to differ in experience with word sequences that typically occur in the domain of job hunting (e.g. *goede contactuele eigenschappen* 'good communication skills', *werving en selectie* 'recruitment and selection'). The groups are not expected to differ systematically in experience with word sequences that are characteristic of news reports (e.g. *de Tweede Kamer* 'the House of Representatives', *op een gegeven moment* 'at a certain point'). For each of these two registers, we selected 35 word sequences and used these as stimuli in two experiments that yield insight into participants' linguistic representations and processing: a completion task and a Voice Onset Time experiment.

In the following section, we discuss the concept of predictive processing in more detail. We describe how prediction in language processing is commonly investigated, focusing on the research design of those studies and the limitations. We then report on the outcomes of our study into variation in predictions and processing speed. The results show that there are meaningful differences to be detected between groups of speakers, and that a small collection of data elicited from the participants themselves can be more informative than general corpus data. The prediction-based effects we observe are shown to be clearly influenced by differences in experience. On the basis of these findings, we argue that it is worthwhile to go beyond amalgamated data whenever prior experiences form a predictor in models of language processing and representation.

4.1.1 Prediction-based processing in language

Context-sensitive prediction is taken to be a fundamental principle of human information processing (Bar 2007; A. Clark 2013). As Bar (2007: 281) puts it, "the brain is continually engaged in generating predictions". These processes have been observed in numerous domains, ranging from the formation of first impressions when meeting a new person (Bar, Neta & Linz 2006), to the gustatory cortices that become active not just when tasting actual food, but also while looking at pictures of food (Simmons, Martin & Barsalou 2005), and the somatosensory cortex that becomes activated in anticipation of tickling, similar to the activation during the actual sensory stimulation (Carlsson, Petrovic, Skare, Petersoon & Ingvar 2000)

In order to generate predictions, the brain "constantly accesses information in memory" (Bar 2007: 288), as predictions rely on associative activation. We extract repeating patterns and statistical regularities from our environment and store them in long-term memory as associations. Whenever we receive new input (from the senses or driven by thought), we seek correspondence between the input and existing representations in memory. We thus activate associated, contextually relevant representations that translate into predictions. So, by generating a prediction, specific regions in the brain that are responsible for processing the type of information that is likely to be encountered are activated. The analogical process can thus assist in the interpretation of subsequent input. Furthermore, it can strengthen and augment the existing representations.

Expectation-based activation comes into play in a wide variety of domains that involve visual and auditory processing (see Bar 2007; A. Clark 2013). Language processing is no exception in this respect (see, for example, Kuperberg & Jaeger 2016). This is in line with the cognitive linguistic framework, which holds that the capacity to acquire and process language is closely linked with fundamental cognitive abilities. In the domain of language processing, prediction entails that language comprehension is dynamic and actively generative. Kuperberg and Jaeger (2016) list an impressive body of studies that provide evidence that readers and listeners anticipate structure and/or semantic information prior to encountering new bottom-up information. People can use multiple types of information –ranging from syntactic, semantic, to phonological, orthographic, and perceptual– within their representation of a given context to predictively pre-activate information and facilitate the processing of new bottom-up inputs.

There are several factors that influence the degree and representational levels to which we predictively pre-activate information (Brothers, Swaab & Traxler 2017; Kuperberg & Jaeger 2016). The extent to which a context is constraining matters (e.g. a context like "The day was breezy so the boy went outside to fly a…" will pre-activate a specific word such as 'kite' to a higher degree than "It was an ordinary day and the boy went outside and saw a…”). Contexts may also differ in the types of representations they constrain for (e.g. they could evoke a specific lexical item, or a semantic schema, like a restaurant script). In addition to that, the comprehender’s goal and the instructions and task demands play a role. Whether you quickly scan, read for pleasure, or carefully process a text, may affect the extent to which you generate predictions. Also the speed at which bottom-up information unfolds is of influence: the faster the rate at which the input is presented, the less opportunity there is to pre-activate information.

The contextually relevant associations that are evoked seem to be pre-activated in a graded manner, through probabilistic prediction. On this account, the mental representations for expected units are activated more than those of less expected items (Roland, Yun, Koenig & Mauner 2012). The expected elements, then, are easier to recognize and process when they appear in subsequent input. When the actual input does not match the expectations, it is more surprising and processing requires more effort.

As Kuperberg and Jaeger (2016) observe, most empirical work has focused on effects of lexical constraint on processing. These studies indicate that a word’s probability in a given context affects processing as reflected in reading times (Fernandez Monsalve, Frank & Vigliocco 2012; McDonald & Shillcock 2003; Roland et al. 2012; Smith & Levy 2013), reaction times (Arnon & Snider 2010; Traxler & Foss 2000), and N400 effects (Brothers, Swaab & Traxler 2015; DeLong, Urbach & Kutas 2005; Frank, Otten, Galli & Vigliocco 2015; Van Berkum, Brown, Zwitserlood, Kooijman & Hagoort 2005). A word’s probability is commonly expressed as cloze probability or surprisal. The former is obtained by presenting participants with a short text fragment and asking them to fill in the blank, naming the most likely word (i.e. a completion task or cloze procedure, W. Taylor 1953). The cloze probability of a particular word in the given context is expressed as the percentage of individuals that complemented the cue with that word (DeLong et al. 2005: 1117). A word’s surprisal is inversely related, through a logarithmic function, to the conditional probability of a word given the sentence so far, as estimated by language models trained on text corpora (Levy 2008). Surprisal thus expresses the extent to which an incoming word deviates from what was predicted.

4.1.2 Usage-based variation in prediction-based processing

The measures that quantify a word’s predictability in studies to date —cloze probabilities and surprisal estimates— are coarse-grained approximations of participants’ experiences. The rationale behind relating processing effort to these scores is that they gauge people’s experiences and resulting predictions. The responses in a completion task are taken to reflect people’s knowledge resulting from prior experiences; the corpora that are used to calculate surprisal are supposed to represent such experiences. However, the cloze probabilities and surprisal estimates are based on amalgamations of data of various speakers, and they are compared to processing data from yet other people. Given that people differ from each other in their experiences, this matter should not be treated lightheartedly. Language acquisition studies have convincingly shown children's language production to be closely linked to their own prior experiences (e.g. Borensztajn, Zuidema & Bod 2009; Dąbrowska & Lieven 2005; Lieven, Salomo & Tomasello 2009). In adults, individual variation in the representation and processing of language has received much less attention.

If we assume that prediction-based processing is strongly informed by people's past experiences, the best way to model processing ease and speed would require a database with all of someone's linguistic experiences. Unfortunately, linguists do not have such databases at their disposal. One way to investigate the relationship between experiences, expectations, and ease of processing is to use groups of speakers who are known to differ in experience with a particular register, and to compare the variation between and within the groups. This can then be contrasted with a register with which the groups' experiences do not differ systematically. Having participants take part in both a task that uncovers their predictions and a task that measures processing speed makes it possible to relate reaction times to participants' own expectations.

A comparison of groups of speakers to reveal usage-based variation appears to be a fruitful approach. Various studies indicate that people with different occupations (Dąbrowska 2008; Gardner, Rothkopf, Lapan & Lafferty 1987; Street & Dąbrowska 2010, 2014), from different social groups (Balota, Cortese, Sergent-Marshall, Spieler & Yap 2004; Caldwell-Harris, Berant & Edelman 2012), or with different amounts of training (Wells, Christiansen, Race, Acheson & MacDonald 2009) vary in the way they process particular words, phrases, or (partially) schematic constructions with which they can be expected to have different amounts of experience. To give an example, Caldwell-Harris and colleagues (2012) compared two groups with different prayer habits: Orthodox Jews and secular Jews. They administered a perceptual identification task in which phrases were briefly flashed on a computer screen, one word immediately after the other. Participants were asked to report the words they saw, in the order in which they saw them. As expected, the two groups did not differ from each other in performance regarding the non-religious stimuli. On the religious phrases, by contrast, Orthodox Jews were found to be more accurate and to show stronger frequency effects than secular Jews. The participants who had greater experience with specific phrases could more easily match the brief, degraded input to a representation in long-term memory, recognize and report it. Note, however, that these studies do not relate the performance on the experimental tasks to any other data from the participants themselves, and, with the exception of Street and Dąbrowska (2010, 2014), the researchers pay little attention to the degree of variation *within* each of the groups of participants.

While we would expect individual differences in experience to affect prediction-based processing, as those predictions are built on prior experience, very little research to date has looked into this. To draw conclusions about the strength of the relationship between predictions and processing effort, and the underlying mental representations, we ought to pay attention to variation across language users. This will, in turn, advance our understanding of the role of experience in language processing and representation and the theoretical status of individual variation.

4.2 Outline of the present research

In this paper, we examine variation between and within three groups of speakers, and we relate the participants' processing data to their own responses on a task that reveals their context-sensitive predictions. Our first research question is: To what extent do differences in amount of experience with a particular register manifest themselves in different expectations about upcoming linguistic elements when faced with word sequences that are characteristic of that register? Our second research question is: To what extent do a participant's own responses in a completion task predict processing speed over and above word predictability measures based on data from other people?

To investigate this, we had three groups of participants —recruiters, job-seekers, and people not (yet) looking for a job— perform two tasks —a completion task and a Voice Onset Time (VOT) task. In both tasks, we used two sets of stimuli: word sequences that typically occur in the domain of job hunting and word sequences that are characteristic of news reports. In the completion task, the participants had to finish a given incomplete phrase (e.g. *goede contactuele …* 'good communication …'), listing all things that came to mind. In the VOT task, the participants were presented with the same cues, followed by a specific target word (e.g. *eigenschappen* 'skills'), which they had to read aloud as quickly as possible. The voice onset times for this target word indicate how quickly it is processed in the given context.

The cue is taken to activate knowledge about the words' co-occurrence patterns based on one's prior experiences. Upon reading the cue, participants thus generate predictions about upcoming linguistic elements. In the completion task, the participants were asked to list these predictions. The purpose of the VOT task is to measure the time it takes to process the target word, in order to examine the extent to which processing is facilitated by the word's predictability. According to prediction-based processing models, the target will be easier to recognize and process when it consists of a word that the participant expected than when it consists of an unexpected word.

As the three groups differ in experience in the domain of job hunting, participants' experiences with these collocations resemble their fellow group members' experiences more than those of the other groups. Consequently, we expect to see on the job ad stimuli that the variation across groups in expectations is larger than the variation within groups. As the groups do not differ systematically in experience with word sequences characteristic of news reports, we expect variation across participants on these stimuli, but no systematic differences between the groups.

Subsequently, we examine to what extent processing speed in the VOT task correlates with participants' expectations as expressed in the completion task. The VOT task yields insight into the degree to which the recognition and pronunciation of the final word of a collocation is influenced not only by the word's own characteristics (i.e. word length and word frequency), but also by the preceding words and the expectations they evoke. By relating the voice onset times to the participant's responses on the completion task, we can investigate, for each participant individually, how a word's contextual expectedness affects processing load. Various studies indicate that word predictability has an effect on reading times, above and beyond the effect of word frequency, possibly even prevailing over word frequency effects (Dambacher, Kliegl, Hofmann & Jacobs 2006; Fernandez Monsalve et al. 2012; Rayner, Ashby, Pollatsek & Reichle 2004; Roland et al. 2012). In these studies, predictability was calculated on the basis of data from people other than the actual participants. As we determine word predictability for each participant individually, we expect our measure to be a significant predictor of processing times, over and above measures based on data from other people.

4.2.1 Participants

122 native speakers of Dutch took part in this study. All of them had completed higher vocational or university education or were in the process of doing so. The participants belong to one of three groups. The first group, labeled *Recruiters*, consists of 40 people (23 female, 17 male) who were working as a recruiter, intermediary, or HR adviser at the time of the experiment. Their ages range from 22 to 64, mean age 36.0 (*SD* = 10.0).

The second group, *Job-seekers*, consists of 40 people (23 female, 17 male) who were selected on the basis of reporting to have read at least three to five job advertisements per week in the three months prior to the experiment, and who never had a job in which they had to read and/or write such ads. Their ages range from 19 to 50, mean age 33.8 (*SD* = 8.6).

The third group, labeled *Inexperienced*, consists of 42 students of Communication and Information Sciences at Tilburg University (28 female, 14 male) who participated for course credit. They were selected on the basis of reporting to have read either no job ads in the past three months, or a few but less than one per week. Furthermore, in the past three years there was not a single month in which they had read 25 job ads or more, and they never had a job in which they had to read and/or write such ads. These participants' ages range from 18 to 26, mean age 20.2 (*SD* = 2.1).

4.2.2 Stimuli

The stimuli consist of 35 word sequences characteristic of job advertisements and 35 word sequences characteristic of news reports. These word sequences were identified by using a Job ad corpus and the Twente News Corpus, and computing log-likelihood following the frequency profiling method of Rayson and Garside (2000). The Job ad corpus was composed by Textkernel, a company specialized in information extraction, web mining and semantic searching and matching in the Human Resources sector. All the job ads retrieved in the year 2011 (slightly over 1.36 million) were compiled, yielding a corpus of 488.41 million tokens. The Twente News Corpus (TwNC) is a corpus of comparable size (460.34 million tokens), comprising a number of national Dutch newspapers, teletext subtitling and autocues of broadcast news shows, and news data downloaded from the Internet (University of Twente, Human Media Interaction n.d.).[20] By means of the frequency profiling method we identified *n*-grams, ranging in length from three to ten words, whose occurrence frequency is statistically higher in one corpus than another, thus appearing to be characteristic of the former (see Kilgarriff 2001). In order to bypass an enormous amount of irrelevant sequences such as *Contract Soort Contract* and _ _ _ _ _, which occur in the headers of the job ads, we applied the criterion that a sequence had to occur at least ten times in one corpus and two times in the other.

[20] The Twente News Corpus represents a fairly broad genre of text, to which the three groups of participants can be presumed to have had similar exposure. The fact that newspapers contain some job ads reflects that participants may have had some exposure to texts of this type even if they are not actively looking for a job or dealing with job ads professionally. The frequency with which they encounter word sequences characteristic of job ads will be much lower, though, than the frequency with which job-seekers and recruiters encounter them. The word sequence "40 uur per week", for example, occurs only 76 times in the entire TwNC.

We selected sequences that met a number of additional requirements. A string had to end in a noun and it had to be comprehensible out of context. We only included *n*-grams that constitute a phrase, with clear syntactic boundaries. Sequences were also chosen in such a way that in the final set of stimuli all content words occur only once.[21] Furthermore, the selected sequences were to cover a range of values on two types of corpus-based measures: sequence frequency and surprisal of the final word in the sequence. With respect to the former, we took into account the frequency with which the sequence occurs in the specialized corpus (i.e. either the Job ad corpus or the News report corpus) as well as a corpus containing generic data, meant to reflect Dutch readers' overall experience, rather than one genre. We used a subset of the Dutch web corpus NLCOW14 (Schäfer & Bildhauer 2012) as a generic corpus. The subset consisted of a random sample of 8 million sentences from NLCOW14, comprising in total 148 million words.

To obtain corpus-based surprisal estimates for the final word in the sequences, language models were trained on the generic corpus. These models were then used to determine the surprisal of the last word of the sequence (henceforth target word). Surprisal was estimated using a 7-gram modified Kneser–Ney algorithm as implemented in SRILM.[22]

The resulting set of stimuli and their frequency and surprisal estimates can be found in Appendices 4.1 and 4.2. The length of the target words, measured in number of letters, ranges from 3 to 17 (News report items *M* = 7.1, *SD* = 3.0, Job ad items *M* = 8.6, *SD* = 3.6). Word length and frequency will be included as factors in the analyses of the VOT data, as they are known to affect processing times.

4.2.3 Procedure

The study consisted of a battery of tasks, administered in one session. Participants were tested individually in a quiet room. At the start of the session they were informed that the purpose of the study was to gain insight into forms of communication in job ads and news reports and that they would be asked to read, complement, and judge short text fragments.

First, participants took part in the completion task in which they had to complete the stimuli of which the final word had been omitted (see Section 4.3.1).

[21] The only exception is the word *goed* 'good', which occurs twice.

[22] SRILM is a toolkit for building and applying statistical language models (Stolcke 2002). Modified Kneser–Ney is a smoothing technique for language models that not only prevents non-zero probabilities for unseen words or *n*-grams, but also attempts to improve the accuracy of the model as a whole (Chen & Goodman 1999). A 7-gram model was used, since the length of the selected word strings did not exceed seven words.

After that, they filled out a questionnaire regarding demographic variables (age, gender, language background) and two short attention-demanding, arithmetic distractor tasks created using the Qualtrics software program. These tasks distracted participants from the word sequences that they had encountered in the completion task and were about to see again in the Voice Onset Time experiment. After that, the VOT experiment started. In this task, participants were shown an incomplete stimulus (i.e. the last word was omitted), and then they saw the final word. They read aloud this target word as quickly as possible (see Section 4.4.1 for more details).

The completion task and the VOT task were administered using E-Prime 2.0 (Psychology Software Tools Inc., Pittsburgh, PA), running on a Windows computer. To record participants' responses, they were fitted with a head-mounted microphone.

4.3 Experiment 1: Completion task

4.3.1 Method

4.3.1.1 Materials

The set of stimulus materials comprised 70 cues, divided over two ITEMTYPES: 35 Job ad cues (see Appendix 4.3) and 35 News report cues (see Appendix 4.4). A cue consists of a test item in which the last word is replaced with three dots (e.g. *goede contactuele …* 'good communication …'). The stimuli were presented in a random order that was the same for all participants, to ensure that any differences between participants' responses are not caused by differences in stimulus order.

4.3.1.2 Procedure

Participants were informed that they were about to see a series of short text fragments. They were instructed to read them out loud and complete them by naming all appropriate complements that immediately come to mind. For this, they were given five seconds per trial. It was emphasized that there is not one correct answer. In order to reduce the risk of chaining (i.e. responding with associations based on a previous response rather than responding to the cue, see McEvoy & Nelson 1982; De Deyne & Storms 2008), participants were shown three examples in which the cue was repeated in every response (e.g. cue: *een kopje …* 'a cup of …', responses: *een kopje koffie, een kopje thee, een kopje suiker* 'a cup of coffee, a cup of tea, a cup of sugar'). In this way, we prompted participants to repeat the cue every time, thus minimizing the risk of chaining.

Participants practiced with five cues that ranged in the degree to which they typically select for a particular complement. They consisted of words unrelated to the experimental items (e.g. *een geruite …* 'a checkered …'). The experimenter stayed in the testing room while the participant completed the practice trials, to make sure the cue was read aloud. The experimenter then left the room for the remainder of the task, which took approximately six minutes.

The first trial was initiated by a button press from the participant. The cues then appeared successively, each cue being shown for 5000 ms in the center of the screen. On each trial, the software recorded a .wav file with a five-second duration, beginning simultaneously with the presentation of the cue.

4.3.1.3 Scoring of responses

All responses were transcribed. The number of responses per cue ranged from zero to four, and varied across items and across participants. Table 4.1 shows the mean number of responses on the two types of stimuli for each of the groups. Mixed ANOVA shows that there is no effect of Group, *F*(2, 119) = 0.18, *p* = .83, meaning that if you consider both item-types together, there are no significant differences across groups in mean number of responses. There is a main effect of Itemtype on the average number of responses, *F*(1, 119) = 38.89, *p* < .001, and an interaction effect between Itemtype and Group, *F*(2, 119) = 16.27, *p* < .001. Pairwise comparisons (using a Šidák adjustment for multiple comparisons) revealed that there is no significant difference between the mean number of responses on the two types of items for Recruiters (*p* = .951), while there is for Job-seekers (*p* < .01) and for Inexperienced participants (*p* < .001). The fact that the latter two groups listed more complements on news report items than they did on job ad items is in line with the fact that these two groups have less experience with Job ad phrases than with News report phrases. Note, however, that a higher number of responses per cue does not necessarily imply a higher degree of similarity to the complements that occur in the specialized corpora: a participant may provide multiple complements that do not occur in the corpus.

Table 4.1 Mean number of responses participants gave per cue; standard deviations between brackets.

	News report cues	Job ad cues
	M (*SD*)	*M* (*SD*)
Recruiters	1.12 (0.25)	1.12 (0.21)
Job-seekers	1.18 (0.31)	1.12 (0.24)
Inexperienced	1.24 (0.28)	1.06 (0.27)

By means of stereotypy points (see Fitzpatrick, Playfoot, Wray & Wright 2015) we quantified how similar each participant's responses are to the complements observed in the specialized corpora. The nominal complements that occurred in the corpus in question were assigned percentages that reflect the relative frequency.[23] The sequence *40 uur per* ('40 hours per'), for example, was always followed by the word *week* ('week') in the Job ad corpus. Therefore, the response *week* was awarded 100 points; all other responses received zero points. In contrast, the sequence *kennis en* ('knowledge and') took seventy-three different nouns as continuations, a few of them occurring relatively often, and most occurring just a couple of times. Each response thus received a corresponding amount of points. For each stimulus, the points obtained by a participant were summed, yielding a stereotypy score ranging from 0 to 100.[24]

4.3.1.4 Statistical analyses

By means of a mixed-effects logistic regression model (Jaeger 2008), we investigated whether there are significant differences across groups of participants and sets of stimuli in the proportion of responses that correspond to a complement observed in the specialized corpora. Mixed-models obviate the necessity of prior averaging over participants and/or items, enabling the researcher to model random subject and item effects (Jaeger 2008). Appendix 4.5 describes our implementation of this statistical technique.

[23] For a given cue [Cue 1], we retrieved all complements in the corpus that consist of a noun that immediately follows the string constituting the cue. This constitutes [Set 1]. For each complement, we determined its token frequency in [Set 1], ignoring any variation in the use of capitals. The sum of all complements' token frequencies is [SumFreq]. A particular complement's stereotypy points were calculated as follows: [complement C_n's token frequency in Set1] / [SumFreq] * 100. If a response in the Completion task corresponded to complement C_n, then that response was assigned C_n's stereotypy points. If a response in the Completion task did not correspond to any complement found in the corpus, then that response was assigned zero stereotypy points.

[24] Stereotypy points are related to, but not the same as, the metrics surprisal and entropy. Entropy quantifies how uncertain the language model is about what will come next. Entropy expresses the uncertainty at position *t* about what will follow; surprisal expresses how unexpected the actually perceived word w_{t+1} is. As Willems et al. (2016: 2507) explain: "if only a small set of words is likely to follow the current context, many words will have (near) zero probability and entropy is low". The word that actually appears in this case may or may not be highly surprising, depending on whether or not it conforms to the prediction. The uncertainty about the upcoming word w_{t+1} does not appear to affect processing of that word w_{t+1} when the effect of surprisal of w_{t+1} has been factored out. It is word w_t that is read more slowly when entropy(t) is higher (Frank 2013; Roark, Bachrach, Cardenas & Pallier 2009).

4.3.2 Results

For each stimulus, participants obtained a stereotypy score that quantifies how similar their responses are to the complements observed in the specialized corpora. Table 4.2 presents the average scores of each of the groups on the two types of stimuli.

Table 4.2 Mean stereotypy scores (on a 0-100 scale); standard deviations between brackets

	News report stimuli	Job ad stimuli
	M (*SD*)	*M* (*SD*)
Recruiters	31.1 (10.9)	42.0 (7.6)
Job-seekers	32.5 (5.5)	34.3 (9.5)
Inexperienced	29.5 (5.5)	18.5 (5.7)

The average scores in Table 4.2 mask variation across participants within each of the groups (as indicated by the standard deviations) and variation across items within each of the two sets of stimuli. Figure 4.1 visualizes for each participant the mean stereotypy score on News report items and the mean stereotypy score on Job ad items. It thus sketches the extent to which scores on the two item types differ, as well as the extent to which participants within a group differ from each other. Figure 4.2 portrays these differences in another manner; it visualizes for each participant the difference in stereotypy scores on the two types of stimuli. The majority of the Recruiters obtained a higher stereotypy score on Job ad stimuli than on News report stimuli, as evidenced by the Recruiters' marks above the zero line. For the vast majority of the Inexperienced participants it is exactly the other way around: their marks are predominantly located below zero. The Job-seekers show a more varied pattern, with some participants scoring higher on Job ad items, some scoring higher on News report items, and some showing hardly any difference between their scores on the two sets of items.

What the figures do not show is the degree of variation across items within each of the two sets of stimuli. The majority of the Recruiters obtained a higher mean stereotypy score on Job ad items than on News report items. Nevertheless, there are several Job ad items on which nearly all Recruiters scored zero (see Appendix 4.3; a group's average stereotypy score of $\leq$10.0 indicates that most group members received zero points on that item) and News report items on which nearly all of them scored 100 (see Appendix 4.4, Recruiters' average scores $\geq$90.0).

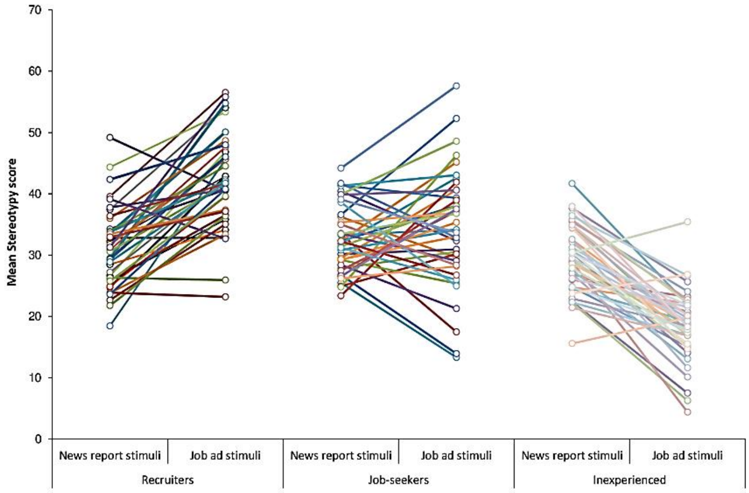


Figure 4.1 Mean stereotypy score on the two types of stimuli for each individual participant.

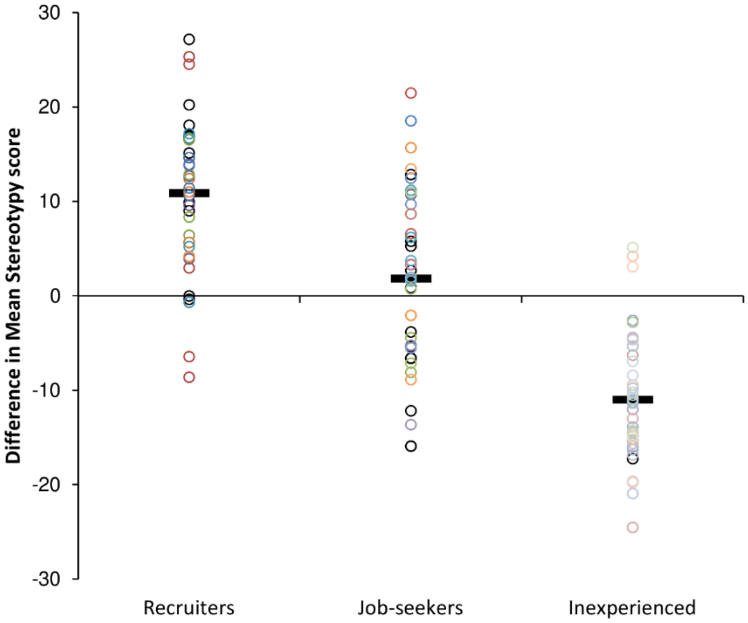


Figure 4.2 The difference between the mean stereotypy score on Job ad stimuli and the mean stereotypy score on News report stimuli for each individual participant; black bars show each group's mean difference. A circle below zero indicates that that participant obtained higher stereotypy scores on News report stimuli than on Job ad stimuli.

By means of a mixed logit-model, we investigated whether there are significant differences between groups and/or item types in the proportion of responses that correspond to a complement observed in the specialized corpora while taking into account variation across items and participants. The model (summarized in Appendix 4.5) yielded four main findings.

First, we compared the groups' performance on News report stimuli. The model showed that there are no significant differences between groups in the proportion of responses that correspond to a complement in the Twente News Corpus. On the Job ad stimuli, by contrast, all groups differ significantly from each other. The Recruiters have a significantly higher proportion of responses to the Job ad stimuli that match a complement in the Job ad corpus than the Jobseekers (β = -0.69, *SE* = 0.17, 99% CI: [-0.11, -0.26]). The Job-seekers, in turn, have a significantly higher proportion of responses to the Job ad stimuli that correspond to a complement in the Job ad corpus than the Inexperienced participants (β = -1.69, *SE* = 0.25, 99% CI: [-2.34, -1.04]).

Subsequently, we examined whether participants' performance on the Job ad stimuli differed from their performance on the News report stimuli. The mixed logit-model revealed that when variation across items and variation across participants are taken into account, the difference in performance on the two types of items does not prove to be significant for any group. However, there were significant interactions. For the Recruiters, the proportion of responses that correspond to a complement in the specialized corpus is slightly higher on the Job ad items than on the News report items, while for the Job-seekers it is the other way around. In this respect, these two groups differ significantly from each other (β = 0.91, *SE* = 0.21, 99% CI: [0.36, 1.46]). For the Inexperienced participants, the proportion of responses that correspond to a complement in the specialized corpus is much higher on the News report items than on the Job ad items. As such, the Inexperienced participants differ significantly from both the Job-seekers (β = 1.23, *SE* = 0.32, 99% CI: [0.38, 2.07]) and the Recruiters (β = 2.14, *SE* = 0.38, 99% CI: [1.15, 3.09]).

4.3.3 Discussion

In this completion task, we investigated participants' knowledge of various multi-word units that typically occur in either news reports or job ads. Participants named the complements that came to mind when reading a cue, and we analyzed to what extent their expectations correspond to the words' co-occurrence patterns in corpus data.

In all three groups, and in both stimulus sets, there is variation across participants and across items in the extent to which responses correspond to corpus data. Still, there is a clear pattern to be observed. On the News Report items, the groups do not differ significantly from each other in the proportion of responses that correspond to a complement observed in the Twente News Corpus. On the Job ad stimuli, by contrast, all groups differ significantly. The Recruiters' responses correspond significantly more often to complements observed in the Job ad corpus than the Job-seekers' responses. The Job-seekers' responses, in turn, correspond significantly more often to a complement in the Job ad corpus than the responses of the Inexperienced participants.

The results indicate that there are differences in participants' knowledge of multi-word units which are related to their degree of experience with these word sequences. This knowledge is the basis for prediction-based processing. Participants' expectations about upcoming linguistic elements, expressed by them in the completion task, are said to affect the effort it takes to process the subsequent input. That is, the subsequent input will be easier to recognize and process when it consists of a word that the participant expected than when it consists of an unexpected word. We investigated whether the data on individual participants' expectations, gathered in the completion task, are a good predictor of processing speed. In a follow-up Voice Onset Time experiment, we presented the cues once again, together with a complement selected by us. Participants were asked to read aloud this target word as quickly as possible. In some cases, this target word had been mentioned by them in the completion task; in other cases, it had not. Participants were expected to process the target word faster – as evidenced by faster voice onset times– if they had mentioned it themselves than if they had not mentioned it.

4.4 Experiment 2: Voice Onset Time task

4.4.1 Method

4.4.1.1 Materials

The set of stimuli comprised the same 70 experimental items as the completion task (35 Job ad word sequences and 35 News report word sequences, described in Section 4.2.2) plus 17 filler items. The fillers were of the same type as the experimental items (i.e. (PREPOSITION) (ARTICLE) ADJECTIVE NOUN) and consisted of words unrelated to these items (e.g. *het prachtige uitzicht* 'the beautiful view'). The stimuli were randomized once. The presentation order was the same for all participants, to ensure that any differences between participants' responses are not caused by differences in stimulus order.

4.4.1.2 Procedure

Each trial began with a fixation mark presented in the center of the screen for a duration ranging from 1200 to 3200 ms (the duration was varied to prevent participants from getting into a fixed rhythm). Then the cue words appeared at the center of the monitor for 1400 ms. A blank screen followed for 750 ms. Subsequently, the target word was presented in blue font in the center of the screen for 1500 ms. Participants were instructed to pronounce the blue word as quickly and accurately as possible. 1500 ms after onset of the target word, a fixation point appeared, marking the start of a new trial.

Participants practiced with eight items meant to range in the degree to which the cue typically selects for a particular complement and in the surprisal of the target word. The practice items consisted of words unrelated to the experimental items (e.g. cue: *een hart van* 'a heart of', target: *steen* 'stone'). The experimenter remained in the testing room while the participant completed the practice trials, to make sure the cue words were not read aloud, as the pronunciation might overlap with the presentation of the target word. The experimenter then left the room for the remainder of the task, which took approximately nine minutes.

The first trial was initiated by a button press from the participant. The stimuli then appeared in succession. After 43 items there was a short break. The very first trial and the one following the break were filler items. On each trial, the software recorded a .wav file with a 1500 ms duration, beginning simultaneously with the presentation of the target word.

All participants performed the task individually in a quiet room. The Inexperienced group was made up of students who were tested in sound-attenuated booths at the university. The Recruiters and Job-seekers were tested in rooms that were quiet, but not as free from distractions as the booths. This appears to have influenced reaction times: the Inexperienced participants responded considerably faster than the other groups (see Section 4.4.2). A by-subject random intercept in the mixed-effects models accounts for structural differences across participants in reaction times.

4.4.1.3 Data preparation and statistical analyses

Mispronunciations were discarded (e.g. stuttering *re- revolutie*, naming part of the cue in addition to the target word *per week*, pronouncing *loge* ('box') as *logé* ('guest') or *lodge* ('lodge')). This resulted in loss of 0.59% of the Job ad data and 1.48% of the News report data. Speech onsets were determined by analyzing the waveforms in Praat (Boersma & Weenink 2015; Kaiser 2013: 144).

Using linear mixed-effects models (Baayen et al,. 2008), we examined whether there are significant differences in VOTs across groups of participants and sets of stimuli, analogous to the analyses of the completion task data. We then investigated to what extent the voice onset times can be predicted by characteristics of the individual items and participants. Our main interest is to examine the relationship between VOTs and three different measures of word predictability. In order to assess this relationship properly, we should take into account possible effects of word length, word frequency, and presentation order, since these factors may influence VOTs. Therefore, we included three sets of factors. The first set concerns features of the target word, regardless of the cue, that are known to affect naming times: the length of the target word and its lemma frequency. The second set relates to artifacts of our experimental design: presentation order and block. The third set consists of the factors of interest to our research question: three different operationalizations of word predictability. The predictor variables are discussed in more detail successively. The details of the modeling procedure are described in Appendix 4.6.

WordLength Longer words take longer to read (e.g. Balota et al. 2004; Kliegl, Grabner, Rolfs & Engbert 2004). Performance on naming tasks has been shown to correlate more with numbers of letters than number of phonemes (Ferrand et al. 2011) or number of syllables (Forster & Chambers 1973). Therefore, we included length in letters of the target word as a predictor.

rLogFreq Word frequency has been shown to affect reading and naming times (Connine, Mullennix, Shernoff & Yelen 1990; Forster & Chambers 1973; Kirsner 1994; McDonald & Shillcock 2003; Roland et al. 2012). It is a proxy for a word's familiarity and probability of occurrence without regard to context. We determined the frequency with which the target words (lemma search) occur in the generic corpus. This corpus comprised a wide range of texts, so as to reflect Dutch readers' overall experience, rather than one genre. The frequency counts were log-transformed. Word length and word frequency were correlated (*r* = -.46), as was to be expected. Frequent words tend to have shorter linguistic forms (Zipf 1935). We residualized word frequency against word length, thus removing the collinearity from the model. The resulting predictor rLogFreq can be used to trace the influence of word frequency on VOTs once word length is taken into account.

PresentationOrder As was reported in the Materials section, the stimuli were presented in a fixed order, the same for all participants. We examined whether there were effects of presentation order (e.g. shorter response times in the course of the experiment because of familiarization with the procedure, or longer response times because of fatigue or boredom), and whether any of the other predictors entered into interaction with PresentationOrder.

Block The experiment consisted of two blocks of stimuli. Between the blocks there was a short break. We checked whether there was an effect of Block.

Various studies indicate that word predictability has an effect on reading and naming times (McDonald & Shillcock 2003; Fernandez Monsalve et al. 2012; Rayner et al. 2004; Roland et al. 2012; Traxler & Foss 2000). Word predictability is commonly expressed by means of corpus-based surprisal estimates or cloze probabilities, using amalgamated data from different people; hardly ever is it determined for participants individually. In our analyses, we compare the following three operationalizations:

GenericSurprisal The surprisal of the target word given the cue, estimated by language models trained on the generic corpus meant to reflect Dutch readers' overall experience (see Section 4.2.2 for more details).[25]

ClozeProbability The percentage of participants that complemented the cue in the completion task preceding the VOT task with the target word. We allowed for small variations, provided that the words shared their morphological stem with the target (e.g. *info* – *informatie*).

TargetMentioned A binary variable that expresses for each participant individually whether or not a target word was expected to occur. For each stimulus, we assessed whether the target had been mentioned by a participant in the completion task. Again, we allowed for small variations, provided that the words shared their stem with the target.

[25] Language models could also be trained on the specialized corpora, instead of the generic corpus. The use of SpecializedSurprisal instead of GenericSurprisal would not yield different outcomes, though; there is no effect of SpecializedSurprisal on VOTs (β = 0.006, *SE* = 0.005, 99% CI: [-0.006, 0.018]).

To give an idea of the number of times the target words were listed in the completion task, Table 4.3 presents the mean percentage of target words mentioned by the participants in each of the groups.

Table 4.3 Mean percentage of targets words that had been mentioned by the participants in the completion task; range between brackets.

	News report stimuli		Job ad stimuli	
	M	(range)	*M*	(range)
Recruiters	31.4	(20.0 – 51.4)	44.0	(20.0 – 60.0)
Job-seekers	31.6	(22.9 – 45.7)	36.6	(14.3 – 62.9)
Inexperienced	28.1	(17.1 – 40.0)	19.3	(2.9 – 40.0)

Finally, we included interactions between rLogFreq and measures of word predictability, as the frequency effect may be weakened, or even absent, when the target is more predictable (Roland et al. 2012).

4.4.2 Results

Table 4.4 presents for each group the mean voice onset time per item type. The Inexperienced participants were generally faster than the other groups, on both types of stimuli. This is likely due to factors irrelevant to our research questions: differences in experimental setting, in experience with participating in experiments, and in age. By-subject random intercepts account for such differences.[26] Of interest to us is the way the VOTs on the two types of items relate to each other, and the extent to which the VOTs can be predicted by measures of word predictability. These topics are discussed successively.

[26] Instead of using the mean VOT of all participants, each participant is assigned a personal intercept value. General differences in reaction times are thus accounted for. A participant that was relatively slow across board will have a higher intercept value than participants that were relatively fast. Apart from that, the participants can resemble or differ from each other in the extent to which their VOTs show effects of the predictor variables. An alternative method of accounting for structural differences across participants in reaction times is to standardize the VOTs. This rules out a by-subject random intercept, since every subject has a mean standardized VOT of zero. The outcomes of a model fitted to standardized VOTs were found not to differ essentially from the outcomes of the model fitted to raw VOTs. Therefore, we only report the latter.

Table 4.4 Mean Voice Onset Times in seconds; standard deviations between brackets.

	News report stimuli	Job ad stimuli
	M (*SD*)	*M* (*SD*)
Recruiters	0.541 (0.14)	0.522 (0.14)
Job-seekers	0.539 (0.15)	0.531 (0.14)
Inexperienced	0.476 (0.12)	0.486 (0.11)

Table 4.4 shows that, on average, the Inexperienced participants responded faster to the News report stimuli than to the Job ad stimuli, while for the other groups it is just the other way around. Figures 4.3 and 4.4 visualize the pattern between the VOTs on the two types of items for each participant individually. For 80% of the Recruiters, the difference in mean VOTs on the two types of stimuli is negative, meaning that they were slightly faster to respond to Job ad stimuli than to News report stimuli. For 62.5% of the Job-seekers and 23.8% of the Inexperienced participants the difference score is below zero. Mixed-effects models fitted to the voice onset times (summarized in Table 4.4 and Figures 4.3 and 4.4) revealed that the Inexperienced participants' data pattern is significantly different from the Recruiters' (β = -0.030, *SE* = 0.007, 99% CI: [-0.048, -0.011]) and the Job-seekers' (β = -0.019, *SE* = 0.005, 99% CI: [-0.034, -0.004]). That is, the fact that the Inexperienced participants tended to be faster on the News report items than on the Job ad items makes them differ significantly from both the Recruiters and the Job-seekers (see Appendix 4.6 for more details).

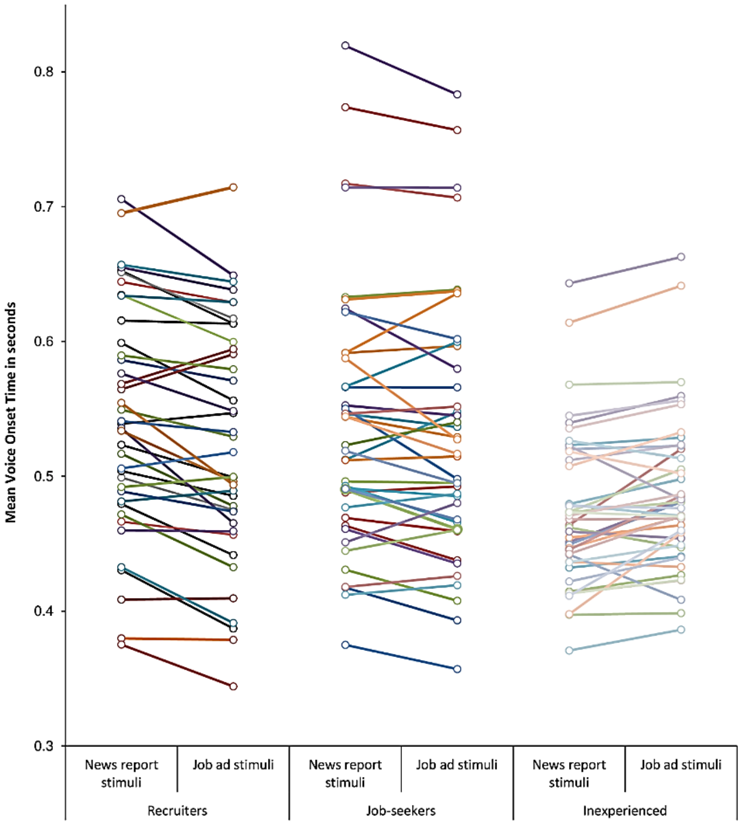


Figure 4.3 Mean Voice Onset Time on the two types of stimuli for each individual participant.

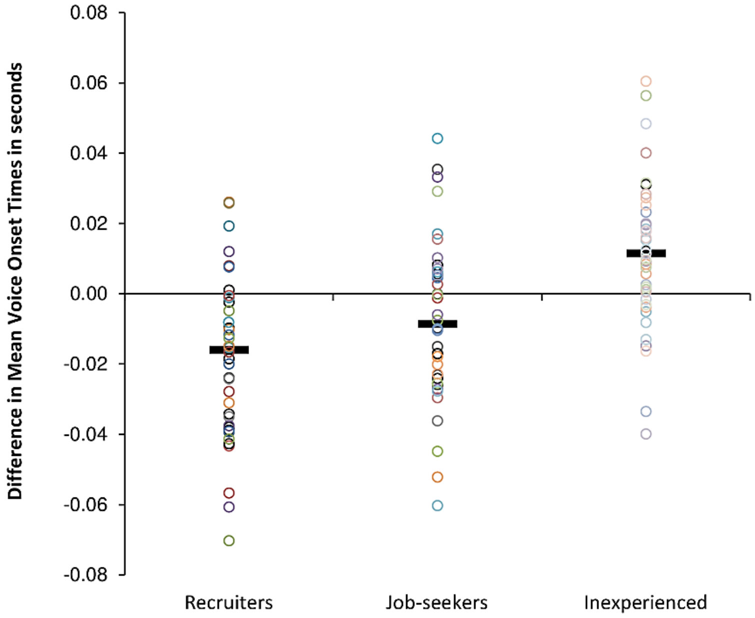


Figure 4.4 The difference between the mean VOT on Job ad stimuli and the mean VOT on News report stimuli for each individual participant; black bars show each group's mean difference. A circle below zero indicates that that participant responded faster on Job ad stimuli than on News report stimuli.

What Figures 4.3 and 4.4 do not show is the degree of variation in VOTs across items within each of the two sets of stimuli. Every mark in Figure 4.3 averages over 35 items that differ from each other in word length, word frequency, and word predictability. By means of mixed-effects models, we examined to what extent these variables predict voice onset times, and whether there are effects of presentation order and block. We incrementally added predictors and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in voice onset times. A detailed description of this model selection procedure can be found in Appendix 4.6. The main outcomes are that the experimental design variable BLOCK and the interaction term PRESENTATIONORDER x BLOCK did not contribute to the fit of the model. The stimulus-related variables WORDLENGTH and rLOGFREQ did contribute. As for the word predictability measures, GENERICSURPRISAL did not improve model fit, but CLOZEPROBABILITY and TARGETMENTIONED did. While the interaction between rLOGFREQ and CLOZEPROBABILITY did not contribute to the fit of the model, the interaction between rLogFreq and TargetMentioned did. None of the interactions of PresentationOrder and the other variables was found to improve goodness-of-fit. The resulting model is summarized in Table 4.5. The variance explained by this model is 60% (R^{2}m = .15, R^{2}c = .60).[27]

Table 4.5 presents the outcomes when *Target not mentioned* is used as the reference condition. The intercept here represents the mean voice onset time when the target had not been mentioned by participants and all of the other predictors take their average value. A predictor's estimated coefficient indicates the change in voice onset times associated with every unit increase in that predictor. The estimated coefficient of rLogFreq, for instance, indicates that, when the target had not been mentioned and all other predictors take their average value, for every unit increase in residualized log frequency, voice onset times are 12 milliseconds faster.

The model shows that ClozeProbability significantly predicted voice onset times: target words with higher cloze probabilities were named faster. In addition to that, there is an effect of TargetMentioned. When participants had mentioned the target word themselves in the completion task, they responded significantly faster than when they had not mentioned the target word (i.e. -0.055).

Lemma frequency (rLogFreq) proved to have an effect when the targets had not been mentioned. When participants had not mentioned the target words in the completion task, higher-frequency words elicited faster responses than lower-frequency words. When the targets had been mentioned, by contrast, word frequency had no effect on VOTs (*B* = -0.001; SE = 0.005; *t* = -0.13; 99% CI = -0.014, 0.012).

Finally, the model shows that while longer words took a bit longer to read, the influence of word length was not pronounced enough to be significant. Presentation order did not have an effect either, indicating that there are no systematic effects of habituation or boredom on response times.

[27] R^{2}m (marginal R^2 coefficient) represents the amount of variance explained by the fixed effects; R^{2}c (conditional R^2 coefficient) is interpreted as variance explained by both fixed and random effects (i.e. the full model) (Johnson 2014).

Table 4.5 Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using *Target not mentioned* as the reference condition.

	Estimate	SE	*t*	99 % CI	
(Intercept)	0.532	0.009	59.86	*0.509, 0.556*	
WordLength	0.012	0.005	2.26	*-0.002, 0.027*	
rLogFreq	-0.012	0.005	-2.58	*-0.024, -0.001*	**
PresentationOrder	0.007	0.005	1.31	*-0.007, 0.020*	
ClozeProbability	-0.025	0.005	-4.64	*-0.039, -0.011*	**
TargetMentioned=yes	-0.055	0.004	-15.11	*-0.065, -0.046*	**
rLogFreq x TargetMentioned=yes	0.011	0.003	4.60	*0.005, 0.018*	**

Note: Significance code: 0.01 '**'

The effects of word frequency (rLogFreq) and TargetMentioned, and the interaction, are visualized in Figure 4.5. All along the frequency range, VOTs were significantly faster when the target had been mentioned by the participants in the preceding completion task. The effect of TargetMentioned is more pronounced for lower-frequency items (the distance between the red and the blue line being larger on the left side than on the right side).

When the targets had not been mentioned, lemma frequency has an effect on VOTs, with more frequent words being responded to faster, as indicated by the descending red line. The effect of frequency is significantly different when the target had been mentioned by participants. In those cases, frequency had no impact.

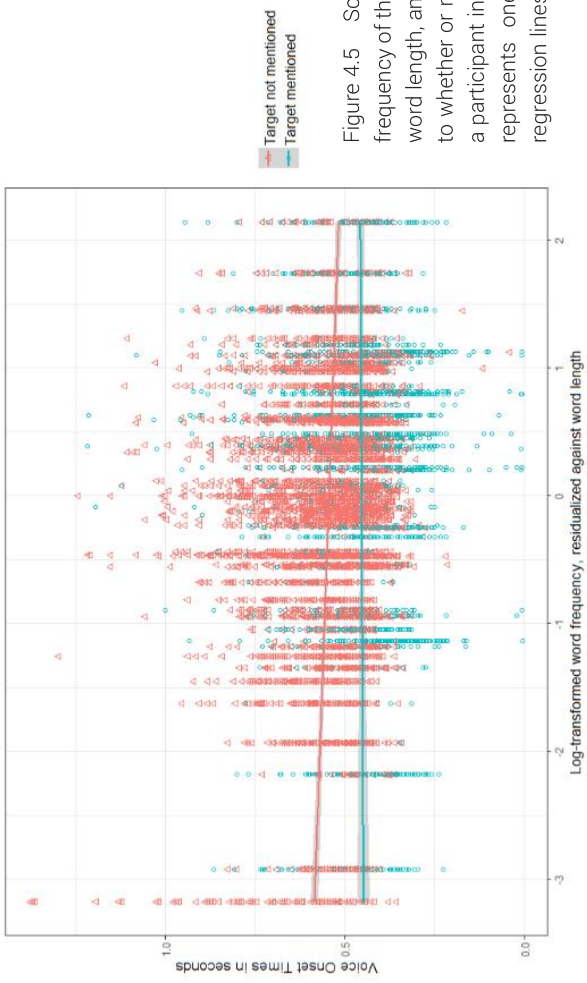


Figure 4.5 Scatterplot of the log-transformed corpus frequency of the target word (lemma), residualized against word length, and the Voice Onset Times, split up according to whether or not the target word had been mentioned by a participant in the preceding completion task. Each circle represents one observation; the lines represent linear regression lines with a 95% confidence interval around it.

4.4.3 Discussion

By means of the Voice Onset Time task, we measured the speed with which participants processed a target word following a given cue. Our analyses revealed that the Inexperienced participants' data pattern was significantly different from the Recruiters' and the Job-seekers': the majority of the Recruiters and the Job-seekers responded faster to the Job ad items than to the News report items, while it was exactly the other way around for the vast majority of the Inexperienced participants.

In all three groups, and in both stimulus sets, there was variation across participants and across items in voice onset times. We examined to what extent this variance could be explained by different measures of word predictability, while accounting for characteristics of the target words (i.e. word length and word frequency) and the experimental design (i.e. presentation order and block). This resulted in five main findings.

First of all, GenericSurprisal, which is the surprisal of the target word given the cue estimated by language models trained on the generic corpus, did not contribute to the fit of the model. In other words, the mental lexicons of our participants could not be adequately assessed by the generic corpus data. It is quite possible that the use of another type of corpus —one that is more representative of the participants' experiences with the word sequences at hand— could result in surprisal estimates that do prove to be a significant predictor of voice onset times. It was not our goal to assess the representativeness of different types of corpora. Studies by Fernandez Monsalve et al. (2012), Frank (2013), and Willems, Frank, Nijhof, Hagoort, and Van den Bosch (2016) offer insight into the ways in which corpus size and composition affect the accuracy of the language models and, consequently, the explanatory power of the surprisal estimates. Still, there may be substantial and systematic differences between corpus-based word probabilities and cloze probabilities, as Smith and Levy (2011) report, and cloze probabilities may be a better predictor of processing effort.

The second finding is that ClozeProbability —a measure of word predictability based on the completion task data of all 122 participants together— significantly predicted voice onset times. Target words with higher cloze probabilities were named faster. Combined, the first and the second finding indicate that general corpus data is too coarse an information source for individual entrenchment, and that the total set of responses in a completion task from the participants themselves forms a better source of information.

Third, our variable TargetMentioned had an effect on voice onset times over and above the effect of ClozeProbability. TargetMentioned is a measure of the predictability of a target for a given participant: if a participant had mentioned this word in the completion task, this person was known to expect it through context-sensitive prediction. Participants were significantly faster to name the target if they had mentioned it themselves in the completion task. This operationalization of predictability differs from those in other studies in that it was determined for each participant individually, instead of being based on amalgamated data from other people. It also differs from priming effects (McNamara 2005; Pickering & Ferreira 2008), which tend to be viewed as non-targeted and rapidly decaying. In our study, participants mentioned various complements in the completion task. Five to fifteen minutes later (depending on a stimulus' order of presentation in each of the two tasks), the target words were presented in the VOT task. These targets were identical, related, or unrelated to the complements named by a participant. The effects of completion task responses on target word processing in a reaction time task are usually not viewed as priming effects, given the relatively long time frame and the conscious and strategic nature of the activation of the words given as a response (see the discussion in Kuperberg & Jaeger 2016: 40; also see Otten & Van Berkum's 2008 distinction between discourse-dependent lexical anticipation and priming).

Both ClozeProbability and TargetMentioned are operationalizations of word predictability. They were found to have complementary explanatory power. ClozeProbability proved to have an effect when the target had not been mentioned by a participant, as well as when the target had been mentioned. In both cases, higher cloze probabilities yielded faster VOTs. This taps into the fact that there are differences in the degree to which the targets presented in the VOT task are expected to occur. A higher degree of expectancy will contribute to faster naming times. The binary variable TargetMentioned does not account for such gradient differences. ClozeProbability, on the other hand, may be a proxy for this; it is likely that targets with higher cloze probabilities are words that are considered more probable than targets with lower cloze probabilities.

Conversely, TargetMentioned explains variance that ClozeProbability does not account for. That is, participants were significantly faster to name the target if they had come up with this word to complete the phrase themselves approximately ten minutes earlier in the completion task. This finding points to actual individual differences and highlights the merits of going beyond amalgamated data. The fact that a measure of a participant's own predictions is a significant predictor of processing speed over and above word predictability measures based on amalgamated data, had not yet been shown in lexical predictive processing research. It does fit in, more generally, with recent studies into the processing of schematic constructions in which individuals' scores from one experiment were found to correlate with their performance on another task (e.g. Misyak, Christiansen & Tomblin 2010; Misyak & Christiansen 2012).

The fourth main finding is that the effect of TargetMentioned on voice onset times was stronger for lower-frequency than for higher-frequency items (the distance between the red and the blue line in Figure 4.5 being larger on the left side than on the right side). The high-frequency target words may be so familiar to the participants that they can process them quickly, regardless of whether or not they had pre-activated them. The processing of low-frequency items, on the other hand, clearly benefits from predictive pre-activation.

Fifth, corpus-based word frequency had no effect on VOTs when the target had been mentioned in the completion task (i.e. *t*=-0.13 for rLogFreq; the blue 'Target mentioned' line in Figure 4.5 is virtually flat). In other words, predictive pre-activation facilitates processing to such an extent that word frequency no longer affects naming latency. When participants had *not* mentioned the target words in the completion task, higher-frequency words elicited faster responses than lower-frequency words (in Table 4.5 rLogFreq is significant (*t*=-2.58); the red 'Target not mentioned' line in Figure 4.5 descends).

4.5 General discussion

Our findings lead to three conclusions. First, there is usage-based variation in the predictions people generate: differences in experiences with a particular register result in different expectations regarding word sequences characteristic of that register, thus pointing to differences in mental representations of language. Second, it is advisable to derive predictability estimates from data obtained from language users closely related to the people participating in the reaction time experiment (i.e. using data from either the participants themselves, or a representative sample of the population in question). Such estimates form a more accurate predictor of processing times than predictability measures based on generic data. Third, we have shown that it is worthwhile to zoom in at the level of individual participants, as an individual's responses in a completion task form a significant predictor of processing times over and above group-based cloze probabilities.

These findings point to a continuity with respect to observations in language acquisition research: the significance of individual differences and the merits of going beyond amalgamated data that have been shown in child language processing, are also observed in adults. Furthermore, our findings are fully in line with theories on context-sensitive prediction in language processing, which hold that predictions are based on one's own prior experiences. Yet in practice, work on predictive processing has paid little attention to variation across speakers in experiences and expectations. Studies investigating the relationship between word predictability and processing speed have always operationalized predictability by means of corpus data or experimental data from people other than those taking part in the reaction time experiments. We empirically demonstrated that such predictability estimates cannot be truly representative for those participants, since people differ from each other in their linguistic experiences and, consequently, in the predictions they generate. While usage-based principles of variation are endorsed more and more (e.g. Barlow & Kemmer 2000; Bybee 2010; Croft 2000; Goldberg 2006; Kristiansen & Dirven 2008; Schmid 2015; Tomasello 2003), often the methodological implications of a usage-based approach are not fully put into practice. In this paper, we show that there is meaningful variation to be detected in prediction and processing, and we demonstrate that it is both feasible and worthwhile to attend to such variation.

We examined variation in experience, predictions, and processing speed by making use of two sets of stimuli, three groups of speakers, and two experimental tasks. Our stimuli consisted of word sequences that typically occur in the domain of job hunting, and word sequences that are characteristic of news reports. The three groups of speakers –viz. recruiters, job-seekers, and people not (yet) looking for a job– differed in experience in the domain of job hunting, while they did not differ systematically in experience with the news report register. All participants took part in two tasks that tap into prediction-based processing. The completion task yielded insight into what participants expect to occur given a particular sequence of words and their previous experiences with such elements. In the Voice Onset Time task we measured the speed with which a specific complement was processed, and we examined the extent to which this is influenced by its predictability for a given participant.

The data from the completion task confirmed our hypotheses regarding the variation within and across groups in the predictions participants generate. On the News Report items, the groups did not differ significantly from each other in how likely participants were to name responses that correspond to the complements observed in the Twente News Corpus. On the Job ad stimuli, by contrast, all groups differed significantly from each other. The Recruiters' responses corresponded significantly more often to complements observed in the Job ad corpus than the Job-seekers' responses. The Job-seekers' responses, in turn, corresponded significantly more often to a complement in the Job ad corpus than the responses of the Inexperienced participants. The responses thus reveal differences in participants' knowledge of multi-word units which are related to their degree of experience with these word sequences.

We then investigated to what extent a participant's own expectations influence the speed with which a specific complement is processed. If the responses in the completion task are an accurate reflection of participants' expectations, and if prediction-based processing models are correct in stating that expectations affect the effort it takes to process subsequent input, then it should take participants less time to process words they had mentioned themselves than words they had not listed. Indeed, whether or not participants had mentioned the target significantly affected voice onset times. What is more, this predictive pre-activation, as captured by the variable TargetMentioned, was found to facilitate processing to such an extent that word frequency could not exert any additional accelerating influence. When participants had mentioned the target word in the completion task, there was no effect of word frequency. This demonstrates the impact of context-sensitive prediction on subsequent processing.

The facilitating effect of expectation-based preparatory activation was strongest for lower-frequency items. This has been observed before, not just with respect to the processing of lexical items (Dambacher et al. 2006; Rayner et al. 2004), but also for other types of constructions (e.g. Wells et al. 2009). It shows that we cannot make general claims about the strength of the effect of predictability on processing speed, as it is modulated by frequency.

Perhaps even more interesting is that the variable TargetMentioned had an effect on voice onset times over and above the effect of ClozeProbability. Participants were significantly faster to name the target if they had mentioned it themselves in the completion task. This shows the importance of going beyond amalgamated data. While this may not come across as surprising, it is seldomly shown or exploited in research on prediction-based processing. Even with a simple binary measure like TargetMentioned, we see that data elicited from an individual participant constitute a powerful predictor for that person’s reaction times. If one were to develop it into a measure that captures gradient differences in word predictability for each participant individually, it might be even more powerful.

Our study has focused on processing of multi-word units. Few linguists will deny there is individual variation in vocabulary inventories. In a usage-based approach to language learning and processing, there is no reason to assume that individual differences are restricted to concrete chunks such as words and phrases. One interesting next step, then, is to investigate to what extent similar differences can be observed for partially schematic or abstract patterns. Some of these constructions (e.g. highly frequent patterns such as transitives) might be expected to show smaller differences, as exposure differs less substantially from person to person. However, recent studies point to individual differences in representations and processing of constructions that were commonly assumed to be shared by all adult native speakers of English (see Kemp, Mitchell & Bryant 2017 on the use of spelling rules for plural nouns and third-person singular present verbs in pseudowords; Street & Dąbrowska 2010, 2014 on passives and quantifiers). Our experimental set-up, which includes multiples tasks executed by the same participants, can also be used to investigate individual variation in processing abstract patterns and constructions.

In conclusion, the results of this study demonstrate the importance of paying attention to usage-based variation in research design and analyses – a methodological refinement that follows from theoretical underpinnings and, in turn, will contribute to a better understanding of language processing and linguistic representations. Not only do groups of speakers differ significantly in their behavior, an individual’s performance in one experiment is shown to have unique additional explanatory power regarding performance in another experiment. This is in line with a conceptualization of language and linguistic representations as inherently dynamic. Variation is ubiquitous, but, crucially, not random. The task that we face when we want to arrive at accurate theories of linguistic representation and processing is to define the factors that determine the degrees of variation between individuals, and this requires going beyond amalgamated data.

Chapter 5 Metalinguistic judgments are psycholinguistic data

5.1 Introduction

Can metalinguistic judgments provide insight into the degree to which linguistic constructions are entrenched in a speaker's mind? A central tenet of usage-based approaches is that language users are sensitive to the distributional properties of the language they encounter and produce. These distributional properties affect the way a linguistic item is represented mentally, which in turn affects the probability that the item will be used, the speed with which it is processed, and the speaker's metalinguistic knowledge regarding its use. From this perspective, degrees of entrenchment of linguistic units can be derived from processing data as well as metalinguistic judgments. On the other hand, processing tasks and judgment tasks may well differ in the processes and knowledge they tap into. Various linguists have voiced the suspicion that entrenchment involves processes which are too deeply embedded for introspection. In this chapter, we present data that contribute to a better understanding of the relationships between metalinguistic judgments, reaction time data, completion task responses, and corpus frequencies, and we test assumptions that follow from a usage-based approach.

5.2 The relationship between metalinguistic judgments and mental representations of language

Usage-based theories posit that mental representations of language emerge from one's experiences with language and general cognitive processes including cross-modal association, categorization, chunking, and analogy (Bybee 2010). These linguistic representations constitute a network of constructions that vary in size and specificity and that are entrenched to different degrees. The more a construction is established as a cognitive routine, the more it is said to be entrenched (Langacker 1987). Usage frequency is a crucial factor in this respect; more experience with a particular construction makes it more strongly entrenched. As a result, the construction can be processed more quickly, fluently, and accurately. Alegre and Gordon (1999), Arnon and Snider (2010), Bybee (2002), and Dąbrowska (2008, 2018), among others, have demonstrated effects of frequency on processing with regard to constructions ranging from morphologically complex words and four-word phrases to (partially) schematic constructions. A question that requires closer investigation is to what extent mental representations are accessible to language users, such that the degree to which linguistic constructions are entrenched in their minds manifests itself not just in processing but also in metalinguistic judgments. In other words, are these degrees of entrenchment part of one's explicit knowledge and can metalinguistic judgments be used to gain insight into entrenchment?

On the one hand, "judgments are the results of linguistic and cognitive processes, by which people attempt to process sentences and then make metalinguistic judgments on the results of those acts of processing (…) Thus, they implicate the same linguistic representations involved in all acts of processing", as Branigan and Pickering (2017: 4) contend. On the other hand, different kinds of linguistic activities –such as reading a text, rapidly making choices in a lexical decision task, completing phrases, reading aloud words, assigning familiarity ratings, making grammaticality judgments– involve different aspects of linguistic representations and may differ in the degree to which they appeal to particular mental representations. Judgments are said to be influenced by knowledge and beliefs (Dąbrowska 2016a) and to reflect decision-making biases (Branigan & Pickering 2017) which are not involved in language processing. What is more, various researchers are concerned that introspections cannot yield accurate insights into subconscious cognitive processes (e.g. Gibbs 2006; Roehr 2008; Stubbs 1993). To assess which aspects are accessible to introspection, it is fruitful to compare metalinguistic judgments with processing data. Such an approach answers Arppe et al.'s (2010:4) call for more multi-methodological research to gain a better understanding of the characteristics and restrictions of each type of evidence.

Prior research has reported correlations between familiarity ratings for various types of lexical units (i.e. words, word pairs, phrases, idioms, and metaphors) and other measures that may provide information on degrees of entrenchment. More specifically, these ratings have proved to be significant predictors of reading times (e.g. Cronk et al. 1993; Juhasz & Rayner 2003; Williams & Morris 2004), performance on lexical decision and speeded naming tasks (e.g. Gernsbacher 1984; Connine et al. 1990; Blasko & Connine 1993; Juhasz et al. 2015), speeded semantic judgment tasks (e.g. Tabossi et al. 2009), and perceptual identification tasks (Caldwell-Harris et al. 2012). While these findings are insightful, they are limited in that the sets of familiarity ratings come from different people than the datasets indicating performance in processing tasks. Multi-method approaches are better able to provide valid insights into task-specific characteristics if they account for individual differences. Speakers differ from each other in their linguistic experiences; their linguistic representations are expected to differ accordingly. If this is the case, we cannot tell whether a discrepancy between familiarity judgments and processing data reflects the fact that different tasks tap into different processes and knowledge, or whether it reflects individual variation in linguistic representations. By having participants who are known to differ in experience with a particular domain of language use, perform a metalinguistic judgment task as well as psycholinguistic processing tasks, we can differentiate between the two.

The goal of this study is two-fold. First, it will reveal to what extent differences in amount of experience with a particular register manifest themselves in different familiarity judgments when faced with word sequences that are characteristic of that register. To this end, three groups of participants –recruiters, job-seekers, and people not (yet) looking for a job– performed a metalinguistic judgment task in which they assigned familiarity ratings to two sets of stimuli – word sequences characteristic of either job ads or news reports. As the three groups differ in experience in the domain of job hunting, they are likely to differ in experience with collocations that are typically used in that domain. According to usage-based theories, these differences in experience lead to differences in mental representations of language. This leads to a testable hypothesis: If familiarity judgments give expression to linguistic representations, the ratings should reflect these differences. That is, the Job ad stimuli ought to be most familiar to the Recruiters and least familiar to the Inexperienced participants.

Subsequently, we examine the relationship between metalinguistic judgments and other types of experimental data. The stimuli that were presented in the judgment task have also been used in two other experiments conducted among the same participants: a completion task and a Voice Onset Time experiment (both described in Chapter 4). By analyzing the judgment data in relation to the participants' completion task responses, their voice onset times, and corpus-based frequencies, we can answer the second research question: To what extent do someone's own data from psycholinguistic processing tasks have explanatory power in predicting familiarity judgments in addition to corpus frequencies? If the different types of tasks tap into the same mental representations, one's performance in the processing tasks should be a significant predictor of one's familiarity ratings. If it does not prove to be a significant predictor, this means that there are substantial differences between the tasks in the information they provide.

5.3 Method

5.3.1 Participants

The same participants that took part in the completion task and the VOT experiment (described in Chapter 4) performed this metalinguistic judgment task. The sample consisted of 122 native speakers of Dutch who belonged to one of three groups: Recruiters, Job-seekers, and Inexperienced participants. Section

4.2.1 provides further details regarding gender, age, educational background, and group membership criteria.

5.3.2 Stimuli

The stimuli were the word sequences also used in the completion task and the VOT experiment. They are described in detail in Chapter 4, Section 4.2.2. The set consists of 35 word strings characteristic of job advertisements and 35 word strings characteristic of news reports, covering a range of phrase frequency values. The stimuli were randomized once for this task. The presentation order was the same for all participants, to ensure that any differences between participants' judgments are not caused by differences in stimulus order.

5.3.3 Judgment task

Participants were asked to rate familiarity using Magnitude Estimation (Bard, Robertson & Sorace 1996). This type of task was also used in the studies reported in Chapters 2 and 3. Instead of using a set judgment scale, participants build their own scale, making as many fine-grained distinctions as they feel appropriate (see Section 2.2.3.2 for more information).

The concept 'familiarity' was defined in the following way: In this task, we ask you to judge how familiar various word combinations are to you. For every word combination, you enter a figure. The more familiar the word combination, the higher the figure. You can think of familiarity in the following ways: you use it often; you hear it often; you read it often (*In deze taak vragen we u te beoordelen hoe vertrouwd verschillende woordcombinaties voor u zijn. Bij elke woordcombinatie vult u een getal in. Hoe vertrouwder de woordcombinatie, hoe hoger het getal. Denk bij vertrouwdheid aan: u gebruikt het vaak; u hoort het vaak; u leest het vaak.*)

5.3.4 Procedure

Participants took part in the familiarity judgment task after they had finished the completion task and the VOT experiment. All word combinations to be judged had occurred in these other tasks. Studies like the one by Lilly (2009) suggest that having seen all stimuli before making any judgments might actually be beneficial (in terms of number of revisions, participants' perception of the reliability of their judgments, and the correlation between judgments and objective scores in studies for which such scores exist).

As in the judgment tasks reported in Chapters 2 and 3, the items were presented in an online questionnaire form (using the Qualtrics software program) and this was also the environment within which the ratings were given. Participants were introduced to the notion of relative ratings through the example of comparing the size of depicted clouds and expressing this relationship in numbers. They were instructed to rate each stimulus relative to the preceding one. A new stimulus was always presented together with the preceding item and the score assigned to it. In a brief practice session, participants then gave familiarity ratings to verb–object combinations (e.g. *een aardappel poffen* 'bake a potato'). They were advised not to start very low, in order to allow for subsequent lower ratings; not to assign negative numbers; and not to set an upper bound a priori.[28]

Before starting the main experiment, participants were informed that the word combinations to be rated had already occurred in the previous tasks. They were asked to assess their familiarity leaving aside the occurrences in this study.

The first six stimuli covered the phrase frequency range of the entire set of items, and the first as well as the seventh stimulus was taken from the middle region of the frequency range, as this may stimulate sensitivity to differences between items with moderate familiarity (Sprouse 2011). Midway, participants were informed that they had completed half of the task and they were offered the opportunity to fill in remarks and questions, just like they were at the end of the task.

5.3.5 Data transformations

For each participant, the ratings were converted to Z-scores to make comparisons of relative ratings possible, just like we did in Chapters 2 and 3. A Z-score of 0 indicates that a particular item is judged by a participant to be of average familiarity compared to the other items. Appendices 5.1 and 5.2 list the mean of the Z-scores of all participants for a given item, and the standard deviation.

5.3.6 Statistical analyses

First, we conducted an analysis of variance and planned contrasts to examine whether there are significant differences in familiarity ratings across groups of participants and sets of stimuli, analogous to the analyses of the completion task data (Section 4.3.2) and the voice onset times (Section 4.4.2).

We then fitted linear mixed-effects models (Baayen et al. 2008), using the LMER function from the lme4 package in R (version 3.3.3; CRAN project; R Core Team, 2017), to the standardized familiarity ratings. We investigated to what extent individual participants' performance on other tasks predicts familiarity ratings, on top of corpus frequencies. To determine this, we included three sets of factors in the model. The first set consists of the corpus-based measures that were also employed in the analyses of the judgment data in Chapters 2 and 3:

[28] The instructions and practice items are available in DataverseNL at https://hdl.handle.net/10411/EL6KZX.

phrase frequency, and lemma frequency of the final word in the phrase.[29] The second set comprises two measures based on individual participants' performance on the preceding experimental tasks involving the same stimuli: the measure TargetMentioned, which is based on participants' completion task responses, and the voice onset times from the VOT experiment. The third set comprises the factors PresentationOrder and Block as artifacts of our experimental design. The predictor variables are discussed in more detail successively; the details of the modeling procedure are described in Appendix 5.3, and the datasets and the scripts are available in DataverseNL at https://hdl.handle.net/10411/EL6KZX.

The variable LogFreqPhrase is the log-transformed frequency with which the phrase as a whole occurs in a subset of the Dutch web corpus NLCOW14 (Schäfer & Bildhauer 2012) – a generic corpus, meant to reflect Dutch readers' overall experience, rather than one genre. The subset consisted of a random sample of 8 million sentences from NLCOW14, comprising in total 148 million words.

LogFreqLemma is the log-transformed lemma frequency of the final word of the phrase in the generic corpus.

The variable TargetMentioned expresses whether or not a participant was known to expect the final word of a stimulus to occur given the preceding words. For each stimulus, we assessed whether the final word (i.e. the target word) had been mentioned by a participant in the completion task. We allowed for small variations, provided that the words shared their morphological stem with the target (e.g. *info* – *informatie*).

The variable VOT is based on the data from the voice onset time experiment. It is the time it took a given participant to start pronouncing the target word as soon as it appeared on screen following the cue (i.e. the stimulus with the final word omitted).

[29] In the analyses of the voice onset times (Chapter 4) we included the factor GenericSurprisal – a measure that is derived from corpus data. It is the surprisal of the final word given the preceding words in the phrase, estimated by language models trained on a generic corpus. Surprisal estimates are commonly used as a measure of word predictability and processing speed. GenericSurprisal is unlikely to be a strong predictor of perceived familiarity of the phrase as a whole once phrase frequency has already been taken into account. We checked whether GenericSurprisal would contribute to explaining variance in familiarity ratings. Adding GenericSurprisal to the model containing LogFreqPhrase did not improve model fit ($\chi^2(1) = 0.38$, $p = .54$). Therefore, we omitted this factor in all subsequent analyses.

Finally, we examined possible effects of PresentationOrder and Block. As was reported in Section 5.3.2, the stimuli were presented in a fixed order, the same for all participants. Halfway there was a short break.

5.4 Results

5.4.1 Variation across groups of participants and sets of stimuli

Figure 5.1 visualizes the mean familiarity rating on the two types of items for each participant individually. Figure 5.2 depicts for each participant the magnitude of the difference between these two scores. These figures are based on standardized ratings. The method of Magnitude Estimation entails that the raw scores from different participants cannot be compared directly, as participants each construct their own scale. Consequently, a score of 50 may represent an average degree of familiarity for one participant, while it expresses a high degree of familiarity for another participant. Once the ratings have been standardized, they can be compared within and between participants.

The fact that the Recruiters display lower scores on the News report phrases than the Inexperienced participants does not mean that these phrases are less familiar to the Recruiters than to the Inexperienced participants in absolute terms. It does mean that the Recruiters consider the Job ad phrases to be more familiar than the News report phrases, while for the Inexperienced participants it is the other way around. The vast majority of the Recruiters (90%) had higher standardized ratings on the Job ad items than on the News report items.[30] The same holds for 77.5% of the Job-seekers and 11.9% of the Inexperienced participants.

An analysis of variance performed on the difference scores depicted in Figure 5.2 showed that there is a significant effect of Group on the difference between participants' mean standardized familiarity rating on Job ad stimuli and their mean standardized familiarity rating on the News report stimuli ($F(2, 71.96) = 74.49$, $p < .001$). Planned contrasts revealed that the difference scores are

[30] There is one notable exception: the Recruiter represented by the turquoise line, whose mean standardized familiarity rating on the News report stimuli amounts to 0.42. We inspected the scores assigned by her. She did not seem to have reversed the scale (i.e. assigning higher ratings to less familiar items), nor did she enter any comments indicating confusion or misunderstanding. The correlation between her standardized ratings and the average of all other participants' standardized ratings is -.17 (Pearson's r). After having analyzed the complete dataset, we also ran the model on the dataset without this participant's data. Excluding her ratings did not alter any of the findings. We decided to keep her scores included, as such a deviant case may be a real characteristic of judgment data.

significantly lower for Inexperienced participants compared to the other groups (t(117.64) = 12.15, $p < .001$, $r = .75$), and that the Job-seekers' difference scores are in turn significantly lower than the Recruiters' (t(77.36) = 2.55, $p < .05$, $r = .28$).[31]

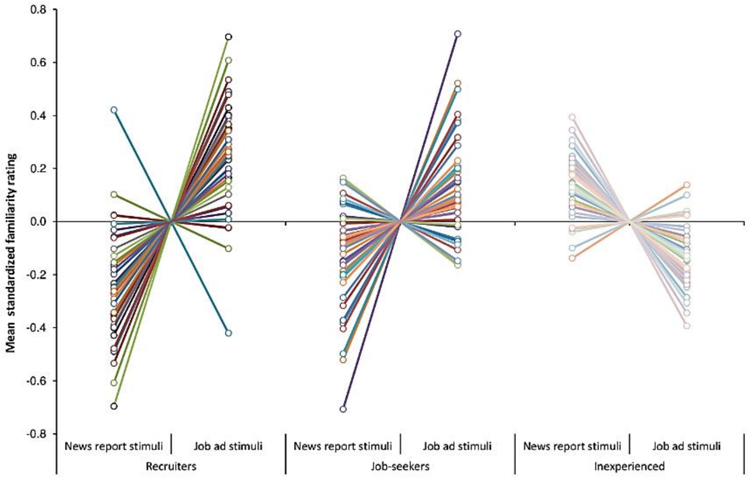


Figure 5.1 Mean standardized familiarity rating on the two types of stimuli for each individual participant.

[31] It was not appropriate to fit a linear mixed-effects model to the standardized familiarity ratings using Group, Itemtype, and their interaction as fixed effects, like we did with the stereotypy scores (Section 4.3.2) and the voice onset times (Section 4.4.2). Such an analysis reveals significant differences across groups in ratings on the News report stimuli, suggesting that these phrases are significantly more familiar to the Inexperienced participants than to the Job-seekers and the Recruiters, while such a conclusion is not justified. Since we had to standardize the ratings, we cannot tell whether these phrases are more familiar to the Inexperienced participants than to the others in absolute terms (cf. Chapter 3, Section 3.5). What we can conclude is that the Inexperienced participants consider the News report stimuli to be more familiar than the Job ad stimuli, while for the Job-seekers and the Recruiters it is the other way around.

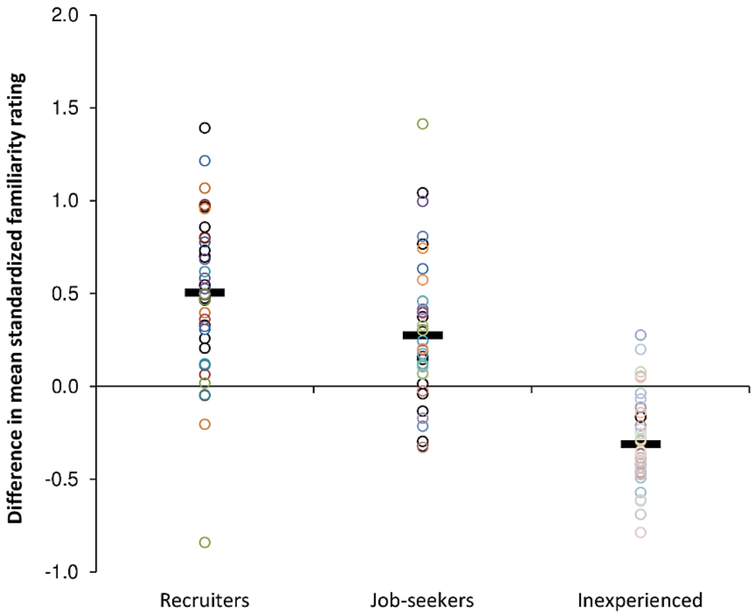


Figure 5.2 The difference between the mean standardized familiarity rating on Job ad stimuli and the mean standardized familiarity rating on the News report stimuli for each individual participant; black bars show each group's mean difference. A circle below zero indicates that that participant assigned higher ratings to News report stimuli than to Job ad stimuli.

5.4.2 Corpus-based frequencies and participant-based psycholinguistic data as predictors of familiarity ratings

Every data point in Figure 5.1 represents the average of the familiarity ratings a participant assigned to 35 stimuli. These items were expected to vary in degree of entrenchment and, consequently, the familiarity ratings were expected to vary too. By means of mixed-effects models, we examined to what extent the variance in ratings can be explained by corpus-based and participant-based measures, and whether there are effects of presentation order and block. We incrementally added predictors and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in ratings. A detailed description of this model selection procedure can be found in Appendix 5.3. The corpus-based measure phrase frequency contributed to the fit of the model; lemma frequency did not, and therefore it was left out. TargetMentioned –a measure of the predictability of a target for a given participant– improved model fit, as did VOT – the time it took the participant to start pronouncing the target word when presented following the cue. Presentation order did not improve model fit, while block did. None of the interactions of block and the other variables was found to improve goodness-of-fit. The interactions of TargetMentioned and the other predictors were included as they did contribute to the fit of the model.

The resulting model is summarized in Tables 5.1 and 5.2. Table 5.1 presents the outcomes when *Target not mentioned* is used as the reference condition. The intercept here represents the mean rating when the target had not been mentioned by participants and all of the other predictors take their average value. A predictor's estimated coefficient indicates the change in ratings associated with every unit increase in that predictor. The estimated coefficient of LogFreqPhrase, for instance, indicates that, when the target had not been mentioned and all other predictors take their average value, for every unit increase in log-transformed phrase frequency, ratings are 0.33 higher. Table 5.2 presents the effects of the predictors when the target *had* been mentioned in the completion task.

Table 5.1 Generalized linear mixed-effects model (family: Gaussian) fitted to the standardized familiarity ratings, using *Target not mentioned* as the reference condition.

	Estimate	*SE*	*t*	99 % CI	
(Intercept)	-0.24	0.07	-3.60	*-0.41, -0.07*	
LogFreqPhrase	**0.33**	**0.05**	**6.65**	***0.20, 0.46***	**
TargetMentioned=yes	**0.45**	**0.03**	**13.87**	***0.37, 0.53***	**
VOT	-0.01	0.01	-1.14	*-0.05, 0.02*	
Block=2	0.23	0.10	2.35	*-0.02, 0.48*	
LogFreqPhrase x TM=yes	**-0.11**	**0.03**	**-3.94**	***-0.18, -0.04***	**
VOT x TM=yes	-0.04	0.02	-1.98	*-0.10, 0.01*	

Note: Significance code: 0.01 '**'

Table 5.2 Generalized linear mixed-effects model (family: Gaussian) fitted to the standardized familiarity ratings, using *Target mentioned* as the reference condition.

	Estimate	*SE*	*t*	99 % CI	
(Intercept)	0.21	0.07	2.97	*0.03, 0.39*	
LogFreqPhrase	**0.22**	**0.05**	**4.26**	***0.08, 0.35***	**
TargetMentioned=no	**-0.45**	**0.03**	**-13.87**	***-0.53, -0.37***	**
VOT	**-0.06**	**0.02**	**-3.36**	***-0.10, -0.01***	**
Block=2	0.23	0.10	2.35	*-0.02, 0.48*	
LogFreqPhrase x TM=no	**0.11**	**0.03**	**3.94**	***0.04, 0.18***	**
VOT x TM=no	0.04	0.02	1.98	*-0.01, 0.10*	

Note: Significance code: 0.01 '**'

First of all, the model shows an effect of TargetMentioned. When participants had mentioned the target word in the completion task, the phrase was given significantly higher familiarity ratings than when the target word had not been mentioned.

In addition, phrase frequency (LogFreqPhrase) proved to have an effect. Higher-frequency phrases were assigned higher familiarity ratings than lower-frequency phrases. This influence of frequency was significantly stronger when the target word had not been mentioned, as is evidenced by the interaction between LogFreqPhrase and TargetMentioned. The effects of TargetMentioned and LogFreqPhrase, and the interaction, are visualized in Figure 5.3. All along the frequency range, ratings were significantly higher when the target had been mentioned by the participants in the preceding completion task. The effect of TargetMentioned is more pronounced for lower-frequency items (the distance

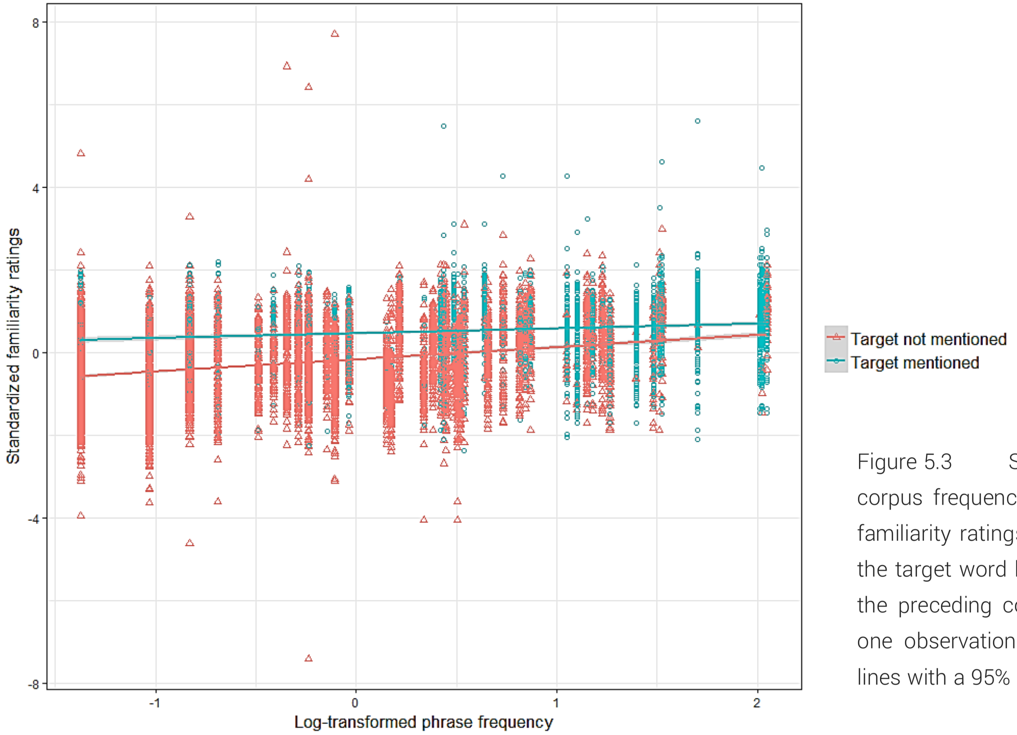


Figure 5.3 Scatterplot of the log-transformed corpus frequency of the phrase, and the standardized familiarity ratings, split up according to whether or not the target word had been mentioned by a participant in the preceding completion task. Each circle represents one observation; the lines represent linear regression lines with a 95% confidence interval around it.

between the red and the blue line being larger on the left side than on the right side). The effect of phrase frequency on familiarity ratings, with more frequent phrases being assigned higher ratings, is indicated by the ascending lines. The phrase frequency effect is significantly stronger when the target had not been mentioned by participants (the red line being steeper than the blue line).

Finally, VOT proved to have an effect when the targets had been mentioned in the completion task. In those cases, it was found that the less time it had taken participants to start pronouncing the target word, the higher the familiarity ratings that were assigned to the phrase. When participants had not mentioned the targets in the completion task, by contrast, the corresponding voice onset times did not predict ratings.

5.5 Discussion

The data presented in this chapter led to two main findings: differences in experiences with a particular register are reflected in the familiarity ratings that participants assign to phrases characteristic of that register; and individual participants' data from a completion task and a Voice Onset Time task are significant predictors of the familiarity ratings they assign to the stimuli. These findings have three important implications. First, they indicate that familiarity judgments and other types of psycholinguistic data tap into the same mental representations of language, and that familiarity ratings form useful data to gain insight into these representations. Second, they provide support for usage-based theories. Third, they add to a growing body of research that points to the significance of individual differences and the merits of going beyond amalgamated data.

In part, these implications follow from the analysis of the differences across groups of participants. Given the differences between the groups in experience with word sequences characteristic of job ads, a usage-based account predicts differences in linguistic representations across groups. If familiarity judgments give expression to linguistic representations, the ratings should reflect these differences, just like the data from the completion task and the Voice Onset Time experiment did. This prediction was borne out. The vast majority of the Recruiters deemed the Job ad phrases to be more familiar than the News report phrases, while for the Inexperienced participants it was the other way around. The Job-seekers are positioned in between the other groups. This pattern resembles the patterns observed in the completion task data (Section 4.3.2) and the voice onset times (Section 4.4.2): most Recruiters obtained a higher stereotypy score and responded faster on Job ad stimuli than on News report stimuli; for the Inexperienced participants it was exactly the other way around; and the Job-seekers took a middle position. Each of these three types of experimental data thus points to usage-based variation in mental representations of multi-word units.

The fact that the judgment data revealed the same patterns we observed in the data from psycholinguistic processing tasks is in line with findings from prior research that related familiarity ratings to processing times (e.g. Blasko & Connine 1993; Caldwell-Harris et al. 2012; Juhasz & Rayner 2003; Juhasz et al. 2015; Tabossi et al. 2009; Williams & Morris 2004). While insightful, these analyses are limited in that they average across participants. In prior research, the judgment tasks and the processing tasks were conducted with different participants. As a result, one cannot distinguish differences across tasks from individual differences which are stable across tasks. By having the same participants perform a metalinguistic judgment task as well as processing tasks and analyzing the data at the level of individual participants, we are able to account for individual differences which are stable across tasks.

We fitted mixed-effects models to the full set of ratings to examine to what extent the variance in ratings can be explained by corpus frequencies and participant-based psycholinguistic data. If the familiarity ratings index the extent and type of previous experience participants have had with the stimulus, then corpus frequencies ought to be a significant predictor of those ratings as they capture variation across items in frequency of occurrence. Corpus-based measures were not expected to explain the variance fully, though, since the corpus is merely a rough approximation of the participants' experiences. Participant-based measures were hypothesized to have additional explanatory power, as they can account for individual differences. If expectations (recorded in the completion task), processing speed (measured in the VOT experiment), and familiarity judgments all reflect the degree to which linguistic units are entrenched in a speaker's mind, then a participant's psycholinguistic responses from the first two tasks can predict that person's familiarity ratings.

As hypothesized, data elicited from an individual participant in other psycholinguistic tasks using the same stimuli constituted a powerful predictor for that person's familiarity judgments. Not surprisingly, corpus-based phrase frequency (LogFreqPhrase) exerted influence, with more frequent phrases being assigned higher ratings. But on top of that, completion task responses and voice onset times were significant predictors. When participants had mentioned the target word in the completion task, they gave significantly higher familiarity ratings to the phrase than when they had not mentioned it. Furthermore, when participants had predicted the final word of a stimulus to occur given the preceding words, corpus-based phrase frequency exerted less influence on their familiarity ratings. In that case, their own voice onset times formed a significant predictor; the faster they had read aloud the target word in the VOT task, the higher the ratings they assigned.

The fact that participants' data from psycholinguistic processing tasks constitute significant predictors of their familiarity judgments when corpus-based frequencies have already been added to the statistical model, is not self-evident. The different tasks each have their own characteristics and restrictions. Voice onset times are more susceptible to noise (e.g. effects of a sneeze or a lapse of attention) than completion task responses or familiarity judgments expressed without time constraints. TargetMentioned, being a binary measure of the predictability of a word, cannot account for gradient differences in entrenchment, while the other two measures can. Metalinguistic judgments, in turn, are said to reflect beliefs and decision-making biases that do not affect online processing, or at least much less so. Nonetheless, there were significant relationships between the different types of data. To be sure, there is variance that is unexplained, and follow-up studies can contribute to a better understanding of the ways in which the tasks differ from each other. Still, the statistical relationships between familiarity ratings and other types of psycholinguistic data as well as corpus frequencies may remove some of the doubts about the usefulness of metalinguistic judgments, at least in investigations of mental representations of multi-word units. The relationships suggest that the different tasks tap into the same linguistic representations. Metalinguistic judgments can be considered psycholinguistic data, just like voice onset times and completion task responses, and they can be as useful as other types of psycholinguistic data to gain insight into linguistic representations.

Our findings showcase the added value of collecting different types of data from the same participants. This practice yields more insight into individual differences and variation across different measures that aim to tap into degree of entrenchment. When researchers work with data sets from different speakers, they are not able to tell to what extent variation is to be ascribed to task-related differences on the one hand, and individual differences in linguistic experience and cognitive abilities on the other. We urge other researchers to conduct multiple tasks among the same speakers, as this will advance our understanding of the cognitive and experiential underpinnings of mental representations of language and the ways in which these representations manifest themselves in various linguistic activities.

Chapter 6

Abstract

Chapters 2 through 5 reported on multi-method studies that involved corpus data as well as offline and online experimental data. What the outcomes imply for theories of mental representations of language is discussed in Chapter 7. The present chapter focuses on the methodological lessons that can be learned from our studies. This chapter highlights the merits of multi-method research in linguistics and may help designing such research. It discusses methodological and practical concerns in the selection of corpus data, metrics to analyze corpus data, stimuli, experimental tasks, and participants, using the studies reported in the previous chapters as case studies.

This chapter is based on:
Verhagen, V., Mos, M., Backus, A. & Schilperoord, J. (submitted). A concise guide to the design of multi-method studies in linguistics: Combining corpus-based measures with offline and online experimental data. *Dutch Journal of Applied Linguistics.* Manuscript submitted for publication.

Chapter 6 A concise guide to the design of multi-method studies in linguistics: Combining corpus-based measures with offline and online experimental data

6.1 Introduction

It is worthwhile to make use of different types of data in linguistic research. This may involve combining data from different sources, using different methods, and integrating quantitative and qualitative approaches, each of which contributes to triangulation (Bryman 2004; Hammersley 2008). Different kinds of data can complement each other, thus yielding a fuller and more precise picture of the phenomena under investigation and more insight into the characteristics and limitations of distinct types of data.

To properly design and conduct a multi-method study requires knowledge from a variety of domains. This chapter presents an overview of the steps to be taken and the decisions to be made. It illustrates this using the studies described in Chapters 2 to 5 as case studies and provides references to useful handbooks and best practices.

In all of our studies, we investigated the relationship between corpus data and experimental data. Corpus data can complement experimental data, for example with respect to ecological validity, contextualization, and scope. Additionally, a comparison of experimental data and corpus data may be used to assess the representativeness of a corpus for particular language users and to test hypotheses regarding the way people process particular linguistic constructions, formulated on the basis of the distributional patterns in a corpus.

Additionally, in Chapters 4 and 5, we examined the relationship between different types of experimental data, by analyzing to what extent performance on one task is a significant predictor of performance on another task. Such analyses enhance our understanding of the extent to which different types of data rely on the same linguistic representations, and how they complement each other. Moreover, we showed the added value of conducting multiple tasks with the same participants – a methodological approach which is, as yet, seldom used in multi-method studies. It makes it possible to distinguish variation across tasks, on the one hand, from variation between participants which is stable across tasks, on the other.

It is not just the combination of different kinds of data that may yield a more complete picture; multiple measurements using the same method can also contribute to more accurate conclusions. In the studies described in Chapters 2 and 3, participants performed the same task twice. We examined the test-retest reliability of metalinguistic judgments and we gained insight into the degree of intra-individual variation relative to inter-individual differences. If the intra-individual variation from one moment to the other reflects the genuine dynamism of linguistic representations, multiple measurements are required to describe this.

Chapters 2 and 3 also yielded insight into the extent to which the outcomes of experiments depend on choices like the type of scale that is used (a 7-point Likert scale or a Magnitude Estimation scale) and whether stimuli are presented in a sentential context or as an isolated word string.

The insights from these studies are used here to discuss the following steps in multi-method research design: selecting corpus data, metrics to analyze corpus data, and stimuli, selecting and designing experimental tasks[32], and selecting participants. Each section concludes with a text box which summarizes the most important considerations, together with useful references.

6.2 Steps in the multi-method approach

6.2.1 Selecting corpus data

A corpus is a collection of texts that can be analyzed for various purposes. It enables you to gain insight into natural language use, obtain large numbers of instances of a linguistic construction (more than is possible via introspection or elicitation) and examine distributional patterns. Such information is of great value, not just in descriptive linguistics; it can serve as a basis to formulate hypotheses on linguistic representations and language processing, and to select items to be used in experiments.

One’s research interests determine which kinds of data are suitable. There is a wide variety of corpus types (see Lüdeling & Kytö 2008 for an overview), differing in terms of size; medium (e.g. written text, transcribed spoken text, audio, video); the time periods that are covered by the texts; the availability of annotations (e.g. part-of-speech tags, lemmatization) and metadata (e.g. text type, information on the writers/speakers); the ways in which the compilers strived to make the corpus representative for a particular language, variety, or register, and balanced such that the proportional sizes of the corpus parts are similar to those in the language, variety, or register. There are no straightforward rules on how to compile or select a good corpus; it greatly depends on your research goal. That is not to say that there are no guidelines (see text box 1).

[32] While acknowledging the merits of other methods, such as ethnography (Levon 2013), interviewing (Schilling 2013), and computational modeling (Pearl 2010), we limit our discussion to experiments.

In our studies, we used existing corpora (i.e. Corpus Gesproken Nederlands, SoNaR, Twente Nieuws Corpus, and NLCOW14), as well as a corpus consisting of Dutch job advertisements that was composed for the purpose of our study. A Dutch job ad corpus did not yet exist. Textkernel, a company that is specialized in information extraction, web mining and semantic searching and matching in the Human Resources sector, created one for us. One of its software modules automatically searches the Internet for new job ads every day. All the job ads retrieved in the year 2011 (slightly over 1.36 million) were compiled, yielding a corpus of 488.41 million words. The past decades have seen developments of software tools and programming languages that make it easier to create corpora and parse and tag the data (see Gries & Newman 2013). It should be noted, though, that a corpus must be constructed carefully for it to be representative.

When building a corpus to be used in multi-method research, it may be possible to collect texts that have been produced (or processed) by the people who will also take part in the experiments. The corpora we used consist of data that our participants had not written themselves, nor had they read all of those texts. Still, the corpora can approximate to their linguistic experiences. It is worthwhile to examine the possibility of compiling a corpus from texts that are produced by the participants themselves. Such a corpus makes it possible to tailor experimental stimuli to a participant's own language use (see, for example, Barking et al. submitted) and to compare corpus-based measures based on either amalgamated data or personal data on how well they correlate with participants' experimental data.

Text box 1. Considerations regarding the selection of corpus data.

What do you want to use the corpus for?

(E.g. to establish characteristics of a particular text type or register, to discover characteristics of particular linguistic constructions, to compare corpus data to experimental data)

What kind of information should the corpus contain?

- Annotations (e.g. part-of-speech tags, lemmatization, phonological annotations)?
- Metadata regarding the texts and/or the authors (e.g. date and place of publication, text type, the writer/speaker's gender, age and nationality)?

Is there an existing corpus that meets your requirements?

Consult overviews such as:

- Lüdeling and Kytö (2008)
- http://corpus.byu.edu/
- http://martinweisser.org/corpora_site/CBLLinks.html
- http://www.inl.nl/taalmaterialen#corpora

Or should you compile one?

Gries and Newman (2013) give useful advice on how to collect and prepare corpus data.

Take into account copyright and privacy issues (see Treadwell 2017 Chapter 3 on collecting Internet content). Check whether your research institute and/or the venue for publishing your work require a research ethics committee to approve of the data collection.

6.2.2 Corpus analysis

Once the corpus data have been selected, you need to decide how to extract information from it. There are an overwhelming number of ways to analyze corpus data and there are ongoing debates (e.g. Bybee 2010; Gries 2012; Schmid & Küchenhoff 2013) as to what metric is most suitable given a particular goal (e.g. do you aim to gauge the predictability of a linguistic unit, its conventionality, its degree of entrenchment out of context, the mutual attraction of lexemes and constructions, the productivity of a construction?). All corpus metrics concern distributions of some kind. Gries and Newman (2013) distinguish three types of distributions of linguistic units: frequencies and dispersion (i.e. how often and where does something occur in a corpus); collocations (i.e. how often do linguistic units occur in close proximity to other linguistic elements); concordances (i.e. how are linguistic units used in their actual contexts, ranging from a few words to whole sentences).

The choice of metrics is motivated by what the corpus-based measures ought to capture, and what the subject of inquiry and the data allow for. If you examine the strength of association between particular words, for instance, you have a choice between unidirectional (either →, or ←) or bidirectional (↔) measures. This choice matters in particular when the strength of association is asymmetric, meaning that one word is more predictive of the other than the other way around (e.g. the word *course* is often preceded by *of*, while there are many different words that tend to follow *of*). When using a corpus-based association measure to predict word-by-word self-paced reading times, a unidirectional measure from left to right (e.g. how predictable is 'president' given the word 'vice', without taking into account to which extent 'president' is predictive of 'vice') may be most in line with the way participants process the language.

It is important to take into consideration that the characteristics of the corpus data and the linguistic constructions of interest may constrain the options. Collostructional analysis (Stefanowitsch & Gries 2003), for instance, is a useful method to analyze the distribution of lexemes in alternating grammatical structures, a common example being the dative alternation. In the case of the dative alternation, collostructional analysis assesses the degree to which particular verbs are attracted to either the prepositional dative (e.g. *she gave the book to him*) or the ditransitive construction (*she gave him the book*). This analysis requires determining the frequency with which the target verb (e.g. *give*) occurs in the target construction (e.g. the prepositional dative), the frequency with which all other verbs occur in the target construction, and the frequency with which the target verb occurs in other constructions. Crucially, in some cases, it may not be possible to define and trace "all other constructions" (this has been called the cell no. 4 problem, see Schmid & Küchenhoff 2013; Bybee 2010 p.98). Apart from the question whether a count of the number of (inflected) verbs can be considered a good proxy for this, you may be faced with the problem that the corpus is not tagged accordingly. In that case, there may be an appropriate tagger available or it might be possible to write a script that can identify and classify relevant parts of speech.

After the selection of metrics, there are usually more decisions to be made. It may be necessary to determine what you consider to be instances of the same construction (e.g. spelling differences like *color* and *colour*, contracted forms like *haven't you* and *have you not*, different forms of the same lemma like *info* and *information* or *has, had,* and *having*). Furthermore, the window around the target item –that is, the amount of context that is taken into account– is to be decided on, as well the possibility to allow for intervening words (e.g. allowing for *een sterk analytisch en probleemoplossend vermogen* 'strong analytical and problem solving skills' to be retrieved when searching for *een sterk analytisch vermogen*).

In addition, it may be important to distinguish between homonyms or different senses of a polysemous word, especially when a query targets a single word. For example, the noun *vermogen* can mean ‘property’, ‘fortune’, ‘capital’, ‘power’, ‘ability’. If some of these uses are irrelevant given the research question, it may be useful to employ word sense disambiguation tools (e.g. WordNet, Princeton University 2010; Agirre & Edmonds 2007).

Finally, when the queries have yielded results, certain transformations may be required. Many metrics are subject to sample-size effects. For example, type-token ratio –a measure of lexical diversity– is known to be affected by text length, with longer texts yielding lower TTR values. If the text segments to be compared in terms of lexical diversity are not of equal sizes, an adjusted score like the mean segmental type-token ratio or the measure of textual lexical diversity can be used (these measures hold either the sample size or TTR constant, see Jarvis 2013). To compare simple frequencies of occurrence of a linguistic construction across (sub)corpora of different sizes, they are normalized as a ratio of occurrences per million words. If the frequencies are to be used as a predictor of experimental data such as processing speed or metalinguistic judgment, it is common to log-transform them, since the relationship between frequency and learning, recognition, and production has been shown to be logarithmic rather than linear (Baayen 2001; Howes & Solomon 1951; Tryk 1986).

To illustrate how research goals and the characteristics of both corpus data and experimental design direct the analysis of the corpus data, the study described in Chapters 4 and 5 can serve as an example. In that study, we were looking for a metric that quantifies the extent to which a word string is characteristic of job advertisements or news reports. We made use of the log-likelihood statistic, following the frequency profiling method of Rayson and Garside (2000). This method enabled us to discover features in the corpora that distinguish one corpus from the other (Kilgarriff 2001). It identifies *n*-grams whose occurrence frequency is statistically higher in one corpus than another, thus appearing to be characteristic of the former. In our comparison of the Job ad corpus and the TwNC, we focused on *n*-grams ranging in length from three to ten words. In order to bypass an enormous amount of irrelevant strings such as *Contract Soort Contract* (‘Contract Type of Contract’), which occur in the headers of the job ads, we applied the criterion that a string had to occur at least ten times in one corpus and two times in the other. As the irrelevant strings do not occur in the TwNC, they are ignored when using this criterion. In this way, we obtained two lists: one containing *n*-grams that appear to be most typical of job ads, and one containing *n*-grams that appear to be most typical of news reports.

We then determined corpus-based string frequency, lemma frequency of the final word, and the degree of unexpectedness of the final word given the preceding words (i.e. surprisal), as these features were expected to affect the way in which the words are processed and judged in our experiments (see Section 6.2.4). Our stimuli are composed of a cue (e.g. *goede contactuele* … 'good communication …') and a target word (e.g. *eigenschappen* 'skills'). The words constituting the cue were presented all at once, and we did not record the speed with which the component parts were processed. Therefore, we did not use corpus-based measures that analyze the internal structure of the cue, and we calculated the surprisal of the target word given the cue as a whole. To obtain surprisal estimates, language models were trained on the generic corpus. A 7-gram model was used, since the length of our word strings did not exceed seven words.

Text box 2. Considerations regarding the analysis of corpus data.

What do you want your corpus-based measure to reveal?
Do your corpus data impose restrictions (e.g. lack of particular kinds of annotations or metadata)?
What type of metrics is most suitable? (Gries 2010a)
(i) Frequencies and dispersion (i.e. how often and where does something occur in a corpus) (Gries 2008)
(ii) Collocations (i.e. how often do linguistic units occur in close proximity to other linguistic elements) (Wiechmann 2008; Divjak 2016)
(iii) Concordances (i.e. how are linguistic elements used in their actual contexts) (Sinclair 1991; Gries 2010b)
Can you make use of an interface in which corpus analysis tools are integrated? For example: - https://portal.clarin.inl.nl/opensonar_whitelab/page/search - http://liwc.wpengine.com/ - http://www.lexically.net/downloads/version5/HTML/, http://lexically.net/wordsmith/step_by_step_Dutch6/index.html?introduction.htm
Which variants of a construction do you want to include or exclude (e.g. spelling differences, contracted forms)?
Are transformations required? If so, which? (Gries 2010a)

6.2.3 Selecting stimuli

In the studies presented in this book, the selection of stimuli for experimental research was based to a large extent on corpus analyses. Analyses of corpus data often play a role in this phase in multi-method research, as such data provide information about characteristics of the linguistic items (e.g. frequency of use, collostructional strength, prototypicality, predictability) that can be used to identify suitable items and predict the way they are processed or rated in experimental tasks.

In the selection of stimuli for experimental research, three main considerations play a role: What is it that the stimuli ought to represent? Which factors ought to be controlled for? How many items are required? Say you want to conduct an eye tracking experiment to examine whether abstract words are processed more slowly than concrete words. Word length and word frequency are known to affect the time it takes to process a word (e.g. Balota, Cortese, Sergent-Marshall, Spieler & Yap 2004), but such effects are not of interest to you. Therefore, you have to control for them. While it may be hard to find sufficient suitable stimuli that do not differ at all in length or frequency, it may be feasible to apply pairwise matching: find a matched control word for every stimulus (i.e. two items that are alike in length and frequency, yet differ in concreteness), or to account for length and frequency effects in the analyses, by including these factors as covariates.

Usually, stimuli constitute a sample, just like participants do. In the case of the concreteness study, the stimuli do not exhaust all possible words in a given language, and the participants do not constitute all speakers of that language. Still, researchers intend to generalize to a population, namely to words beyond the items included in the stimuli set, and to language users beyond the actual people participating. To obtain replicable results that generalize across participant as well as stimulus samples, both sample sizes need to be sufficiently large (see Westfall, Kenny & Judd 2014 for practical tools and guidance). It is important to realize that a suboptimal sample of stimuli can hardly be compensated for by recruiting more participants.

In our selection of 35 Job ad stimuli and 35 News report stimuli for the study reported on in Chapters 4 and 5, we used the following criteria. The words strings had to end in a noun and they had to be comprehensible out of context. We only included *n*-grams that constitute a phrase (more specifically, a noun phrase, a prepositional phrase, or an adjective phrase). It is not clear whether ‘phrasehood’ could have an effect (cf. Arnon & Cohen Priva 2013; Tily et al. 2009). We decided to use only phrases, because we presented the items as isolated word strings. Processing and judging a word string in isolation is less natural for non-constituents than for constituents (compare, for example, *as far as I* to *at the very last moment*).

The word strings were to cover a range of values on two types of corpus-based measures: string frequency and surprisal of the final word in the string, as we aimed to investigate how these variables affect processing and familiarity ratings. Finally, strings were chosen in such a way that in the final set of stimuli all content words occur only once. The stimuli vary in terms of length and frequency of the final word; we included those factors in the analyses.

Text box 3. Considerations regarding the selection of stimuli.

What are the categories or ranges that your stimuli ought to cover? See Cohen (1990) on 'less is more' concerning dependent and independent variables.

Which variables will you control for in stimulus selection and/or analyses?

- Examples of variables to manipulate or control for: (Baayen 2010)
- Databases with norms and ratings for the purpose of stimulus selection: (Keuleers & Balota 2015 for an overview; Juhasz, Lai & Woodcock 2015; McRae, Cree, Seidenberg & McNorgan 2005; Nelson, McEvoy & Schreiber 1998)

Are artificial stimuli required?

Nonwords

- Example of research using Dutch words and nonwords: (Keuleers, Diependaele & Brysbaert 2010)
- Examples of nonword generators and databases: Wuggy (Keuleers & Brysbaert 2010; http://crr.ugent.be/programs-data/wuggy), the English Lexicon Project (Balota et al. 2007, http://elexicon.wustl.edu/), ARC Nonword Database (Rastle, Harrington & Coltheart 2002)

Artificial language

- Examples: Misyak and Christiansen (2012); Van den Bemd, Mos, Alishahi, and Shayan (2014)

What is the appropriate sample size? (Westfall et al. 2014)

6.2.4 Selecting and designing experimental tasks

There is a whole range of experimental methods to choose from, differing on several dimensions. They vary in terms of the modality in which stimuli are presented or produced (e.g. visual, auditory) and whether they involve language comprehension, production, and/or judgment. Furthermore, methods can be characterized as more online or more offline, the former meaning that the method taps into real-time aspects of language processing (e.g. eye tracking), the latter that it assesses the outcomes of this process (e.g. how participants interpret or judge a sentence). Experiments can also be classified as yielding quantitative and/or qualitative data. In addition, experiments differ as to how natural the stimuli are, whether participants are to do something they normally do not do, and how natural the circumstances are in which the task is performed. This has implications for the ecological validity of the study. Self-paced reading (SPR) using a word-by-word moving window, for example, can be considered fairly unnatural. Usually, a sentence is not presented to us one word at a time, and during natural reading we can backtrack and look ahead, while in SPR this is not possible.

Since each type of experiment has its advantages and disadvantages, there is clear added value in combining different types. They can complement each other and thus offer a more complete picture of the subject of investigation (for more elaborate considerations see Arppe, Gilquin, Glynn, Hilpert & Zeschel 2010; Schönefeld 2011). If possible, conduct different experiments with the same participants. When you compare data of tasks conducted with different participants, you are faced with individual differences as well as task-specific contributions to the effects you want to investigate (Connine, Mullennix, Shernoff & Yelen 1990; Chapters 4 and 5).

When participants are to perform a series of tasks, researchers should consider the order carefully, taking into account possible carry-over effects (Myers, Well & Lorch 2010). If they intend to measure effects of surprisal on processing speed, participants should not have seen the target items before. By contrast, if participants are to rate the stimuli, it might actually be beneficial if they have seen all stimuli before making any judgments. In that case, participants have been found to make fewer and smaller revisions when offered the opportunity, and their ratings most closely matched objective scores in studies for which such scores existed (Lilly 2009).

In our last study (Chapters 4 and 5), participants performed a series of tasks in one session. In the completion task, they read out loud the stimuli of which the final word had been omitted (e.g. the cue *een vliegende …* 'a flying …') and completed them by naming all appropriate complements that came to mind within five seconds. After this first task, participants were given a questionnaire regarding demographic variables and two short attention-demanding arithmetic tasks. These small tasks distracted them from the word strings that they had encountered in the completion task and were about to see again in the voice onset time (VOT) experiment. At the same time, the tasks prepared them for the judgment task, illustrating the method of magnitude estimation by which participants build their own scale (Bard, Robertson & Sorace 1996).

In the VOT experiment, the participants were presented with the same cues, followed by a particular target word (e.g. *start* 'start'), which they had to read aloud as quickly as possible. We measured the time it took to recognize and pronounce a particular word following a given word sequence. The 70 stimuli were mixed with 17 filler items, which were of the same type as the experimental items (i.e. (preposition) (article) adjective noun), but consisted of words unrelated to these items (e.g. *het prachtige uitzicht* 'the beautiful view'). The fillers were new to the participants and made the task a bit more varied. The fixation mark that signaled the start of a new trial was displayed on the screen with varying durations, to prevent participants from getting into a fixed rhythm.

Finally, in the judgment task, participants rated how familiar the 70 word strings were to them using Magnitude Estimation (ME). We opted for a judgment task in which participants constructed their own scale, rather than offering a binary or Likert-type fixed set of rating options. In the study presented in Chapter 3, we compared familiarity judgments expressed on a 7-point Likert scale and a Magnitude Estimation scale. The two types of ratings did not differ significantly; both showed a significant effect of phrase frequency (i.e. higher phrase frequency led to higher familiarity ratings, as expected); and there was a near perfect Time1–Time2 correlation of the mean ratings in all experimental conditions. Still, there are some differences worth considering when selecting a particular scale. A Likert scale, unlike a ME scale, makes it possible to determine whether participants consider the majority of items to be familiar (or unfamiliar) and to examine whether all stimuli received a higher rating in a second rating session, provided that there is no ceiling effect preventing increased familiarity to be expressed for certain items. On the other hand, there is a risk that the number of response options on a Likert scale does not match well with the degrees of familiarity as perceived by participants. When offered the opportunity to distinguish more than seven degrees of familiarity, 83.3% of the participants in our study did so. If a Likert scale is opted for, it would be advisable to carefully consider the number of response options.

Prior to the start of an experimental task, participants practiced with items that consisted of words unrelated to the experimental stimuli. For each task, we randomized the stimuli once and kept the presentation order the same for all participants. The reason for this is that we were interested in variation across participants and we wanted to ensure that any differences between participants' responses were not caused by differences in stimulus order. We examined whether there were effects of presentation order (such as shorter response times in the course of the experiment because of familiarization with the procedure, or longer response times because of fatigue or boredom) by including the factor presentation order as a predictor in the statistical analyses.

Another decision to be made is whether the stimuli are presented in isolation or embedded in a context. This may affect the generalizability of the results. In natural language use, linguistic items are encountered in a context and this context can influence the way in which words are interpreted, processed, and responded to. In our first study (Chapter 2), we investigated potential effects of context on familiarity judgments. Participants rated 44 prepositional phrases which were presented as isolated word strings and embedded in a sentence that constituted a prototypical context, resembling the contexts in which the phrase occurred most frequently in the Corpus of Spoken Dutch (CGN). Adding such a context did not yield significantly different judgments. Whether this also holds for other kinds of contexts, varying in size and prototypicality, is yet to be investigated. For possible effects of context in other types of experiments, see studies like those by Burmester et al. (2014), Camblin et al. (2007), and Griffin and Bock (1998) and overviews like Kuperberg and Jaeger's (2016).

Lastly, it is worth considering the insights that can be gained by having participants perform the same experiment twice. While the merits of combining different types of experiments (e.g. Arppe et al. 2010) and replicating a particular study (Andringa & Godfroid 2019; Koole & Lakens 2012) are acknowledged and promoted these days, there seems to be less attention for the value of multiple measurements using the same method, stimuli, and participants. Different kinds of tasks may complement each other (Schönefeld 2011); replications reveal to what extent findings hold when new participants and/or new stimuli are used (Schmidt 2009). The added value of conducting an experiment twice with the same group of subjects is that it leads to a better understanding of the dynamism of mental representations within one language user. Multiple measurements may reveal that the picture that emerged from a single measurement is incomplete and oversimplified.

If participants are to perform an experiment twice, the researcher will have to decide on the test-retest interval. In our first two studies, in which we examined intra-individual variation in metalinguistic judgments across time, participants completed the task twice within a period of one to three weeks. They knew in advance that the experiment involved two test sessions, but not that they would be doing the same task twice. We opted for this time frame as it would be short enough for the construct being tested not to have changed much (i.e. perceived degree of familiarity of phrases that may be used in everyday life, based on at least 18 years of linguistic experiences), yet long enough for the participants not to be able to recall the exact scores they assigned to each of the 88 (Chapter 2) or 79 (Chapter 3) stimuli. The experimental design allowed us to examine variation in judgments within participants from one moment to other, and to compare this to variation between participants.

Text box 4. Considerations regarding the selection and design of experimental tasks.

What do you want your experimental data to reveal?

- Overview of judgment tasks: Schütze and Sprouse (2013)
- Overview of experimental paradigms: Blom and Unsworth (2010), Kaiser (2013)
- Consider the added value of collecting different types of data (Schönefeld, 2011)

What type of design is most useful for your research questions?
Fully crossed, counterbalanced, stimuli-within-condition, participants-within-condition, both-within-condition (Westfall et al. 2014)

In what order should you present your tasks and stimuli?
Randomized, counterbalanced, kept constant (Myers et al. 2010: 412)

Will you embed the stimuli in a context (see Burmester et al. 2014; Camblin et al. 2007; Kuperberg & Jaeger 2016; Chapter 2)? If so, consider the position of the stimulus in the context and the extent to which this may affect processing (e.g. wrap-up effects in eye tracking).

Will you include breaks during tasks? If so, consider starting with a filler item right after a break in an online task, just in case the participants are not yet fully focused upon recommencement.

Will you conduct an experiment twice with the same participants? See Chapter 2 for the merits of doing this and considerations regarding the test-retest interval.

Draw up a protocol, to ensure that all participants are tested in the same way, and conduct a pilot study to detect mistakes, bugs, and lack of clarity.

6.2.5 Selecting participants

In the process of selecting participants, the first step is to define the population you intend to generalize to. After the characteristics of the population have been specified, participants can be selected in such a way that they constitute a representative sample that lends itself to generalizations. Note that while the majority of the publications in behavioral sciences is based on data from Western undergraduates, this subpopulation is among the least representative groups of participants for generalizing about human behavior in general (Henrich, Heine & Norenzayan 2010). Recruiting other types of participants requires some more creativity. If you are looking for a sample of Dutch adults and data collection can take place outside a lab, you could think of visiting the Driver and Vehicle Licensing Agency –an agency that is visited by people from all strata of society– and inviting the visitors who are waiting there for someone taking the driving test to take part in your research. This sample of participants is a more faithful reflection of Dutch society than a group of undergraduates.

Subsequently, consideration should be given to the information that is required to adequately categorize and characterize participants, and to account for their performance in experiments. The challenge is to strike a balance; on the one hand, researchers should not purposelessly collect all kinds of background information, on the other hand, they should not simply assume that particular people will, or will not, differ from each other in certain respects. For example, kindergartners in Wassenaar –an affluent suburb of The Hague, populated by a large expatriate community– need not all be monolingual speakers of Dutch. This should be verified, for example by means of a questionnaire. By gathering information regarding relevant variables, it is possible to determine whether the participants meet the requirements (e.g. monolingual Dutch) and to find out whether they differ from each other on confounding variables (e.g. working memory capacity; see text box 5 for examples of other variables and methods to assess them). If the latter proves to be the case, these data can be included in the analyses as covariates. Alternatively, matching can be used to ensure that the groups to be compared no longer differ in those respects. In pairwise matching, each participant in the first group has a match in the second group in terms of working memory capacity. In groupwise matching, participants are included in the second group if their working memory capacity falls within the first group’s WMC-range.

Another decision that calls for deliberation is the number of participants to be included. If the analyses involve statistical tests, too small a sample size can make the study underpowered. This may give rise to several problems, such as a reduced chance of finding true effects. It is therefore highly recommended to conduct a power analysis and determine the appropriate sample size (see Westfall et al. 2014). If there is a risk of participants dropping out (e.g. in studies comprising multiple test sessions) or data loss (e.g. in eye tracking research), more data can be collected as a measure of precaution.

In our last study (Chapters 4 and 5), we were looking for participants who belonged to one of three groups: recruiters, job-seekers, and people not (yet) looking for a job. We chose these groups as they were expected to differ in experience in the domain of job hunting. We strived to make sure that they did not differ systematically in other respects, such as mother tongue and education level.

We contacted recruitment agencies, as well as HR managers at universities and colleges in Noord-Brabant (a province in the south of the Netherlands), and we conducted our study with 40 people with professional experience with job ads. To recruit job-seekers, we got in touch with the Dutch employee insurance agency (UWV). They sent out a message to approximately 1200 people who were registered as having completed higher vocational or university education, informing them about the opportunity to take part in our study voluntarily. 47 of them were able to participate at that time. To find people who did not (yet) have any experience with job ads, we invited the first-year bachelor and premaster students of Communication and Information Sciences at Tilburg University who were native speakers of Dutch to participate for course credit. 72 students completed the battery of tasks.

As part of the experimental tasks, participants filled out a questionnaire regarding demographic variables. After the last experiment, participants were presented with three questions about their experience with job ad texts. We asked them how many job ads they had read in the past three months (encompassing both thorough reading and scanning); how many months there had been in the past three years in which they read at least 25 job ads per month; whether they ever had a job in which they regularly read or wrote job ads, and if so, for how long. 42 students qualified as inexperienced participants, as they reported to have read either no job ads in the past three months, or a few but less than one per week. Furthermore, in the past three years there was not a single month in which they had read 25 job ads or more, and they never had a job in which they had to read and/or write such ads. As for the job-seekers, we selected those who reported to have read at least three to five job advertisements per week in the three months prior to the experiment, and who never had a job in which they had to read and/or write such ads. This left us with 40 job-seekers.

We made sure that all participants were native speakers of Dutch who had spent most of their youth in the Netherlands, and who had completed higher vocational or university education or were in the process of doing so. The groups could not be matched in terms of age. It was not feasible to find sufficient people who were of the same age as the recruiters and the job-seekers, yet did not have any experience with job ads, or highly-educated recruiters and job-seekers who were of the same age as the inexperienced participants. The difference in age may play a role in the voice onset time experiment, since older adults have slower reaction times. This is not an insurmountable issue, as it is possible to account for structural differences across participants in reaction times in the statistical analyses of the experimental data (for example, by means of a by-subject random intercept in mixed-effects models).

Text box 5. Considerations regarding the selection of participants.

What is/are the population(s)?

See e.g Unsworth and Blom (2010) on comparing L1 and L2 children and adults; Paradis (2010) on comparing typically-developing children and children with specific language impairment.

<table>
<tr><td colspan="2">What sampling technique should you use?
(E.g. convenience, random, stratified, snowball [Buchstaller & Khattab 2013])</td></tr>
<tr><td colspan="2">How, and how much, do participants differ on relevant variables? These data can be used in pairwise or groupwise matching (Paradis 2010), or they can be included as co-varying factors in the analyses. Examples of potentially relevant variables and methods to assess them are listed successively. Only collect those data for which you have a sound theoretical basis to suspect that they play a role.</td></tr>
<tr><td>Demographic
gender, age, ethnicity, regional background, educational background, occupation, socio-economic status</td><td>- Questionnaire (e.g. Gutiérrez-Clellen & Kreiter 2003; Tanner, Inoue & Osterhout 2014)</td></tr>
<tr><td>Linguistic
mother tongue(s), other languages spoken, age of arrival in an L2 environment, learner motivation, amount of exposure to a particular language, language proficiency, vocabulary size</td><td>- Questionnaire (e.g. Tanner et al. 2014)
- Author Recognition Test (Stanovich & West 1989) to assess reading experience
- Language proficiency test (e.g. Hulstijn 2010)
- Peabody Picture Vocabulary Test (Dunn & Dunn 1997)
- SILS Vocabulary Subtest (Zachary 1994)</td></tr>
<tr><td>Cognitive
working memory capacity, nonverbal intelligence, statistical learning ability, need for cognition</td><td>- Verbal working memory span, assessed for example in Daneman and Carpenter’s (1980) or Waters and Caplan’s (1996) reading span task
- Phonological short-term memory, assessed for example in a nonword repetition task (Gupta 2003)
- Short-term memory span, assessed for example in an auditory Forward Digit Span task (Wechsler 1981)
- See Olsthoorn, Andringa and Hulstijn (2014) on working memory capacity tests for natives and nonnative speakers</td></tr>
</table>

	- Raven’s Advanced Progressive Matrices test (Raven, Raven & Court 1998) - the Culture Fair Intelligence Test (Cattell 1971) - (Non)adjacent-dependencies artificial grammar learning (Gómez 2002; Misyak & Christiansen 2012) - Need for Cognition scale (Cacioppo, Petty & Kao 1984)
What is the appropriate sample size (Westfall et al. 2014)? Take into account chances of drop-out and data loss.	
Research ethics Assess potential risks and benefits to your participants and ways to guarantee confidentiality and anonymity. Make sure you comply with the current Code of Conduct and regulations of your research institute. Check whether your institute and/or the venue for publishing your work require a research ethics committee’s approval. Acquire informed consent from participants, and debrief them once they have completed the tasks (see Eckert 2013, and see Treadwell 2017 Chapter 3 in the case of conducting research on the Internet).	

6.2.6 Concluding remarks

The preceding sections discussed methodological and practical considerations in the selection of corpus data, metrics to analyze corpus data, stimuli, experimental tasks, and participants. While these steps cover a significant part of the research process, they are by no means all-encompassing. A well-designed study also entails careful consideration of data management and transparency with respect to the goals and the decisions that are made (see text box 6). Furthermore, multi-method research yields multifaceted datasets that require statistical analyses that do justice to the nature of the data. Which kinds of tests are most suitable depends on the types of data (e.g. continuous or categorical; the number of levels of a categorical variable; whether a variable is nested in other variables, as in the case of children grouped in classes, which are in turn nested in schools) and the research questions (see, for example, Chapters 14 to 16 in Podesva & Sharma 2013; see Chapters 2 to 5 for the analyses employed in our studies).

Multi-method research is often more challenging than mono-method research as regards the operationalization of constructs across methods and data analysis. There is great merit in taking up these challenges, as it leads to more robust evidence, a more complete picture of the subject of research, and a better understanding of the characteristics and limitations of different methods. Various examples of studies in linguistics that successfully combine different types of data can be found in Schönefeld (2011) and Arppe et al. (2011). What our studies (Chapters 2 to 5) add to that, is that they showcase the added value of conducting multiple experiments with the same participants and having participants perform the same task twice. The present chapter discussed these possibilities as part of considerations in the design of multi-method studies. Hopefully, the coming years will see a rise in multi-method studies and constructive debates about the relationships between different types of data and the cognitive representations and processes they tap into.

Text box 6. Considerations regarding research project management.

Preregistration and registered reports

Preregistration entails that you register your research questions and analysis plan prior to data collection (https://cos.io/prereg/). If you opt for a registered report, your research proposal is reviewed prior to data collection (https://cos.io/rr/). If accepted, your findings will be published irrespective of the outcome, provided that you followed the registered plan or provide justification for deviating from it. Pre-registration and registered reports help to get nonsignificant findings published and they foster fair research practices and replications (Nosek & Lakens 2013).

Data management

How will the data be stored, for how long, and who will have access to which parts? Consider repositories like https://dataverse.nl/, https://www.surfdrive.nl/.

Chapter 7 Discussion

The studies presented in this dissertation aim to contribute to theories about the mental representation of language, by examining variation in entrenchment of multi-word units. Cognitive Linguistics takes a usage-based perspective, meaning that mental representations of language are taken to emerge from, and are continuously shaped by, language use. The more frequently a speaker encounters and uses a particular linguistic structure, the more the mental representation of this structure becomes entrenched. As a result, it can be activated and processed more quickly, which, in turn, increases the probability that this form is used to express the given message, making this construction even more entrenched. Conversely, extended periods of disuse weaken the representation (Langacker 1987: 59). Thus, in a usage-based approach, linguistic representations are inherently dynamic: they change over time, and they may differ from one person to another, depending on differences in usage.

While variation naturally follows from a usage-based perspective, surprisingly little is known about how variable mental representations of language are. Corpora, which take a prominent position in usage-based research, contain usage data, but they are nearly always an amalgamation of data from many different people. They may yield insight into the degree of conventionalization of a linguistic construction in a community, but they cannot directly reveal degrees of entrenchment of mental representations (Schmid 2010). As for experimental data, most researchers analyse and report these data at the level of aggregated scores, thus masking individual differences. Often, they relate such data to scores like cloze probabilities and corpus-based measures, which are based on amalgamated data from yet other people. The merit of such an approach is that it has demonstrated robust correlations between frequency of occurrence of linguistic constructions and behavioral indices of cognitive routinization. The drawback is it that disregards inter- and intra-individual variation, while insight into variation is a prerequisite for a veridical description of mental representations of language. In studies of child language acquisition, children's productions have been shown to be closely linked to their own prior experiences (e.g. Borensztajn, Zuidema & Bod 2009; Dąbrowska & Lieven 2005; Lieven, Salomo & Tomasello 2009). In adult native speakers, by contrast, individual differences in representation and processing of language have received much less attention, even though such differences are to be expected on theoretical grounds.

Recent years have seen a growth of interest in social and behavioral sciences in the analysis of individual differences and the limitations of aggregated data as representative of individuals' knowledge and behavior (e.g. Isakov et al. 2016;

Nurius & Macy 2008; Seto et al. 2016; Vindras et al. 2012; von Eye & Bogat 2006; von Eye et al. 2006). Also in the field of cognitive linguistics, this topic is given a more prominent position (e.g. Andringa & Dąbrowska 2019; Barking et al. submitted; Barlow 2013; Dąbrowska 2018; Zimmerer et al. 2011). My dissertation contributes to this strand of research. The studies reported here examined variation between and within participants in metalinguistic judgments on and processing of multi-word sequences. They investigated the variation to be detected and the extent to which this variation can be considered meaningful. In the present chapter, I summarize the main findings and consider the theoretical implications.

7.1 Summary of the findings and their implications

Before analyzing variation between and within individual participants, checks were performed to ascertain that the data display usage-based variation on a more coarse-grained level. The comparison of data from three groups of participants – recruiters, job-seekers, and people not (yet) looking for a job– constituted a first test of usage-based principles. This comparison yielded clear evidence for the relationship between amount of experience with a particular register and (i) the expectations people generate about upcoming words given the initial part of a word string characteristic of that register (Chapter 4); (ii) the speed with which they process such word strings (Chapter 4); and (iii) how familiar they consider these word strings to be (Chapter 5). The results indicate that there are systematic differences in participants' knowledge and processing of multi-word units which are related to their degree of experience with these word sequences. This forms empirical support for a hypothesis that follows from usage-based theories of linguistic knowledge and language processing. As the three groups differ in experience in the domain of job hunting, participants' experiences with these collocations resemble their fellow group members' experiences more than those of the other groups. Consequently, on the job ad stimuli, the variation across groups in expectations, reaction times, and familiarity judgments is hypothesized to be larger than the variation within groups. This hypothesis was supported by the data.

Another finding that was to be expected from a usage-based perspective, is that corpus-based word and phrase frequency correlated with familiarity ratings and reaction times. Higher-frequency phrases were assigned higher familiarity ratings (Chapter 5 as well as Chapters 2 and 3), and when the target word had not been mentioned in the completion task, higher-frequency words elicited faster responses than lower-frequency words (Chapter 4).

The next step involved an investigation of individual differences. It is not just groups of speakers that differ systematically and meaningfully; also at the level of individual speakers, there are meaningful differences to be detected. No two speakers are identical in their linguistic experiences. Usage-based theories thus predict individual differences in entrenchment of linguistic constructions. Indeed, there turned out to be significant relationships between data elicited from an individual participant in different types of psycholinguistic tasks using the same stimuli. A measure of a participant's own predictions (recorded in the completion task) was a significant predictor of that participant's processing speed (measured in the voice onset time experiment). Furthermore, individual participants' data from the completion task and the VOT task were significant predictors of the familiarity ratings they assigned to the stimuli. What is more, these participant-based measures were significant predictors on top of measures based on amalgamated data of different people (i.e. corpus-based frequencies and surprisal; cloze probabilities). In other words, participant-based measures proved to have unique additional explanatory power. This demonstrates the existence of systematic, measurable inter-individual variation in behavioral indices of cognitive routinization.

In addition, there is evidence of intra-individual variation which, too, points to the dynamic character of mental representations of language. A test-retest design provided insight in this kind of variation and illustrates the added value of multiple measurements. In the studies reported in Chapters 2 and 3, participants performed a familiarity judgment task twice within a couple of weeks. The ratings correlated significantly with corpus-based frequencies, just like familiarity ratings in Chapter 5 did. Moreover, analyzing the data at the level of aggregate ratings revealed a near perfect Time1–Time2 correlation. While these findings are interesting, additional insights are obtained by zooming in. None of the participants were as stable in their ratings as the aggregated ratings are; and no single item elicited stable ratings from all participants. The intra-individual variability of metalinguistic judgments could be interpreted as a lack of precision in expressing degree of familiarity. However, it is worth considering an alternative interpretation, namely that the variability in judgments reflects the variability of mental representations of language – at least of multi-word units; whether this may hold for other types of constructions as well is discussed in Section 7.2. Most psycholinguistic tasks that try to tap into the degree of entrenchment of a linguistic unit in the mind of a speaker, express this in a single value (e.g. a rating, a reaction time). However, if cognitive representations can best be viewed as more, or less, densely populated clouds of exemplars that vary in strength depending on frequency and recency of use, a single score yields an incomplete picture. Therefore, I not only advocate attending to variation across participants, I also urge cognitive linguists to carry out multiple measurements per participant.

In sum, the studies yielded support for hypotheses that follow from a usage-based approach, and the insights that were gained upon exploring inter- and intra-individual variation tie in well with such an approach. The findings are an encouragement to flesh out and refine the usage-based framework by studying variation. In what follows, I sketch three compelling directions for future research that build on the work presented in this dissertation. I propose to further develop participant-based measures, to follow participants in the course of a few weeks or months, and to examine (partially) schematic constructions in addition to lexically specific ones. These developments will advance our understanding of the relations between patterns in aggregated data and individual speakers' mental representations.

7.2 Suggestions for future research

The findings presented in this dissertation invite us to further develop theories on the relationship between language in the community and language in the mind, and to formulate and test hypotheses on this matter. Various researchers have drawn attention to the distinction between community-level phenomena and cognitive phenomena in individual speakers (e.g. Backus 2015; Dąbrowska 2016b; Schmid 2015). It follows that patterns observed in aggregated data cannot simply be assumed to be represented as such in the minds of all speakers (e.g. Dąbrowska 2008, 2010; Schmid 2010; Schmid & Mantlik 2015; Zimmerer et al. 2011). For one thing, linguistic constructions are not used by all speakers to the same extent. Furthermore, the mental representations which are activated while processing a particular utterance may differ from one speaker to another, and within one speaker from one moment to another. Usage-based models of language would benefit from a better understanding of individual differences and the relationships between patterns in aggregated data and individual speakers' mental representations. To this end, more insight is required into the ways in which, and the extent to which, individuals differ from each other.

My studies contributed to this by examining variation in metalinguistic knowledge and processing of multi-word units, and contrasting group-based measures and participant-based measures as predictors of judgment data and reaction times. Aggregated data, if sufficiently representative, proved to be significant predictors of participants' familiarity ratings and voice onset times. More specifically, cloze probabilities –a measure of word predictability based on the completion task data of all 122 participants together– significantly predicted voice onset times (Chapter 4), and corpus-based phrase frequency was a significant predictor of familiarity judgments (Chapters 2, 3, and 5). These relationships are to be expected given the large body of work showing correlations between corpus frequencies and data from psycholinguistic experiments, yet more work is needed to fully understand the nature of the relationships between frequency of occurrence based on aggregated data and individual participants' performance on psycholinguistic tasks.

To serve as a predictor of speakers' processing speed and perceived degree of familiarity, aggregated data have to be representative of the participants' experiences with the word sequences at hand. This may seem obvious, yet it is far from easy to specify what exactly qualifies as representative. The content of the corpus is of crucial importance, even more so than corpus size (see, for example, Blom et al. 2012). The better the content reflects the linguistic experiences of the participants, the better the corpus-based measures predict their performance in experiments, as has been shown in studies that assessed how well different types of corpus data predict performance of a given group of participants (e.g. Blom et al. 2012) or how well one type of corpus data predicts performance of different groups of participants (e.g. Gardner et al. 1987). It may be beneficial in this respect to work with more specific definitions of speech communities. This will allow for a more precise analysis of effects of group membership, and more insights into the extent to which entrenchment is determined by usage frequency and by other factors (such as cognitive abilities). As a researcher interested in the use of particular constructions, it would be good to ascertain whether these constructions are characteristic of particular speech communities, define the groups in question, and specify ways to determine to what extent a speaker can be considered a member of a certain community. In Chapters 4 and 5, I started with a priori constructed groups. On the basis of a set of criteria regarding work experience in the field of HR and reported number of job ads read within particular time frames, participants were selected and classified as belonging to one of three groups (i.e. recruiters, job-seekers, and people not (yet) looking for a job). Comparisons of psycholinguistic data revealed that the three groups differ systematically in participants' knowledge and processing of multi-word units characteristic of job ads, while they do not differ significantly on word sequences characteristic of news reports. Subsequently, I analyzed individual variation within groups. Instead of defining groups a priori, people can be classified in a data-driven manner, by identifying user profiles in the data on reported experience (for example, by means of cluster analysis, configural frequency analyses, latent class analysis, or latent class mixture models (von Eye et al. 2004, 2006). This may yield different demarcation lines and result in more (or less) fine-grained groupings than a priori classifications. Subsequently, analyses of the psycholinguistic data can reveal to what extent aggregate level statements hold for the different subgroups and for individuals.

My studies indicated that there are significant relationships between aggregated data and individual speakers' judgment data and reaction times, while at the same time aggregated data mask meaningful variation and do not suffice if the goal is to describe mental representations of language. The variable TargetMentioned, based on data elicited from an individual participant, had an effect on voice onset times over and above the effects of target word corpus frequency and cloze probability (Chapter 4). Participants were significantly faster to name the target if they had mentioned it themselves in the completion task. Similarly, in Chapter 5, the participant-based variables TargetMentioned and VOT were significant predictors of familiarity judgments when corpus-based phrase frequency had already been added to the statistical model. These findings demonstrate inter-individual variation which is stable across different types of experimental tasks. In addition, Chapters 2 and 3 provided insight into intra-individual variation in metalinguistic judgments. Both inter- and intra-individual variation are characteristic of mental representations of language. Group-based measures –such as corpus frequencies and cloze probabilities based on completion task data from different people– are insufficient if the goal is to gain a better understanding of the dynamic nature of linguistic representations.

To this end, additional measures are to be developed and compared, and I hope that my studies can serve as an example and incite researchers to build on this. To obtain data on individual participants' context-sensitive expectations, I conducted the completion task (Chapter 4). Participants listed all appropriate complements that came to mind within five seconds. The responses yield insight in predictions, reflecting what is top of mind for a participant at a given point in time. While surprisal estimates based on corpus data allow for gradual differences across complements in predictability, the set of scores is static. Participants' responses to a completion task, by contrast, may vary from one person to another, and from one moment to another. As such, they are better able to do justice to the variability in associations and ease of activation.

To be able to use the completion task responses as a predictor of processing speed and perceived degree of familiarity, the data were converted into a score that indicates for each participant individually whether the target word had been mentioned or not. This variable, TargetMentioned, proved to be a valuable measure, being a significant predictor of voice onset times in a subsequent naming task, as well as familiarity judgments. However, as a binary variable, it does not account for gradient differences in the degree to which words are expected to occur. It is worth exploring whether the order in which complements are listed by a participant, the number of complements, and perhaps the time it took the participant to come up with a complement provide useful information in this respect. Furthermore, additional tasks can be used to address the fact that certain complements which are not listed by a participant may be familiar to that person nonetheless. It would be interesting to have participants perform the completion task twice and examine the variability in answers from one moment to another. For lower-frequency items, such variation is likely to be larger and responses may be more strongly influenced by recent experiences (e.g. the complements participants listed for the stimulus *een internationale speler van* 'an international player of/from' were often related to the soccer match broadcasted the night before). In addition, stimuli could be provided with more linguistic context, which may affect the saliency and ease of activation of particular complements. These are just a few suggestions as to how the potential of participant-based data can be explored. My studies form a first step, illustrating that it is possible to construct participant-based measures and worthwhile to do so, as they offer new opportunities to gain insight into inter- and intra-individual variation.

Apart from developing new measures, the outcomes of my studies also form an invitation to extend the research questions and experimental designs to other kinds of linguistic constructions and developments over time. The approach I adopted proved to be an effective way to test hypotheses that follow from a usage-based approach. That is, a comparison of data from groups of speakers to reveal usage-based variation was shown to be a fruitful approach in Chapters 4 and 5. The practice of conducting multiple tasks among the same participants yields more insight into variation across speakers on the one hand, and variation across different measures that aim to tap into degree of entrenchment on the other. Conducting the same task twice with the same participants, as in Chapters 2 and 3, allows a better understanding of the relative degrees of inter- and intra-individual variation. This dissertation serves as a proof of concept; the approach can now be applied more extensively and hypotheses can be formulated and tested on a more fine-grained level.

For one thing, it would be interesting to follow participants in the course of a few weeks or months, extending the test-retest design. This can provide additional insights into the effects of usage frequency on processing speed and perceived degree of familiarity. It is clear by now that frequency is a key factor. What is not so clear, is to what extent recency of use matters; whether it makes a difference whether you used a linguistic item once or twice that day; and whether this works differently for low-frequency items compared to high-frequency ones. Such questions can be addressed by tracking people who are in various stages of acquiring a particular jargon (e.g. job ad jargon as a recruiter, political jargon and officialese as a city councilor, statistics jargon as an undergraduate) or language (be it an artificial language like Klingon, or a natural language – a promising project, in this respect, is Marie Barking's, which examines developments in usage and processing of transferred constructions by German students acquiring Dutch). Suppose a study involves participants who just started to work as a recruiter and participants who have had this occupation for various amounts of time. It may be possible to keep track of the texts they read and write during working hours in the course of the investigation. This allows for personal corpora that provide information on individual participants' experiences with a given linguistic construction over time. Experimental data (e.g. on expectations, processing speed, phonological reduction, perceived degree of familiarity) collected in a repeated measures design can be analyzed to identify developmental pathways over time (Nurius & Macy 2008; von Eye et al. 2004) and examine the relationships with developments in usage frequency. This will yield insight into the process of entrenchment, rather than just the product – something usage-based theories will benefit from.

Another direction for future research, which I have not yet explored, is to examine other kinds of linguistic constructions. This dissertation focused on multi-word units – a type of construction that lends itself well to the investigation of usage-based variation. The next step is to examine whether similar degrees of variation can be observed for (partially) schematic constructions. From a usage-based perspective, multi-word units are not essentially different from morphemes, words, partially schematic constructions, or fully abstract schemas. They vary in size and specificity, but they are in essence all form-meaning pairings whose linguistic representation emerges from experience with language together with general cognitive skills such as categorization, schematization, and chunking (Barlow & Kemmer 2000; Bybee 2010). However, it is as yet an open question whether (partially) schematic constructions will display similar degrees of variation as lexically specific constructions.

On the one hand, mental representations of (partially) schematic constructions, too, are dynamic in nature, as they emerge from linguistic experiences. On the other hand, schematic constructions tend to have a more general meaning, a wider range of usage contexts, and a higher frequency of occurrence than lexically specific constructions, which may result in less inter- and intra-individual variability. While the specific instances people encounter and use may differ, the commonalities in meaning and structure could enable them to arrive at similar abstract representations, and differences in the token frequency of the schematic construction may be relatively small. Even when there are individual differences in amount of experience with a schematic construction, the usage frequencies may be so high across the board that there are no detectable differences in processing speed across participants (i.e. a ceiling effect, see for example Street & Dąbrowska 2014; Wells, Christiansen, Race, Acheson & MacDonald 2009).

Although there are reasons to expect schematic constructions to display smaller degrees of individual variation, there are reports of variation among adult native speakers in knowledge of a range of (partially) schematic constructions (e.g. postmodifying PP, object cleft, object relative, simple locatives with the quantifier 'every', possessive locatives with the quantifier 'every', Polish dative inflections) and these individual differences seem to be associated with differences in linguistic experience (Dąbrowska 2008, 2018; Wells et al. 2009). These findings call for more research into usage-based variation in mental representations of (partially) schematic constructions, to examine whether it is similar to what is shown in this dissertation with respect to multi-word units.

Importantly, native speakers do not just differ in the language they encounter and produce, they also differ in cognitive abilities, such as language analytic ability, statistical learning ability, fluid intelligence, and cognitive motivation (Dąbrowska 2018; Misyak & Christiansen 2012). Both linguistic experiences and cognitive abilities appear to influence the process of schematization and speakers' knowledge of grammatical constructions. There are indications that this does not hold for collocational knowledge in the same manner. Dąbrowska (2018) found participants' knowledge of collocations to depend primarily on experience-related factors (print exposure and years spent in full-time education); it did not depend on language analytic ability and non-verbal IQ, while performance on grammatical comprehension and receptive vocabulary tasks did. Such findings do not conflict with usage-based models of linguistic knowledge, but they do call for a refinement of the theoretical framework regarding the ways in which mental representations of language emerge and develop. While representations of words, multi-word units, and grammatical patterns can still be construed as constructions that emerge from linguistic experience together with general cognitive skills, they may differ in the extent to which they rely on various cognitive and experiential factors. Research that aims to advance our understanding of the contributions of these factors must pay attention to individual differences. It seems plausible that highly educated speakers, because of their cognitive abilities as well as their social backgrounds, tend to receive input that makes schematic constructions more salient to them. Furthermore, the tasks they face at school and at work likely invite them to use these schematic constructions more frequently than less-educated speakers. In order to gain more insight into effects of amount and type of experience on the one hand, and cognitive abilities on the other, future studies could make use of (partially) schematic constructions which are characteristic of particular registers, and participants who vary in experience with these registers. Dąbrowska (2018) used print exposure (measured by means of an author recognition test) and years spent in full-time education as variables that reflect linguistic experience. These are rather coarse-grained measures of the amount and type of experience with passives, postmodifying PPs, and universal quantifier constructions. In future studies, it may be possible to identify constructions which are characteristic of specific registers, and obtain more fine-grained information on participants’ degrees of experience with these registers. Experimental data on knowledge and processing of these constructions can then be analyzed in relation to amount of experience and cognitive abilities. I hope this dissertation contributes to this research agenda by demonstrating that it is feasible and valuable to attend to inter- and intra-individual variation and by sparking linguists’ enthusiasm for such an approach.

References

Agirre, E. & Edmonds, P. (2007). *Word sense disambiguation: Algorithms and applications*. Retrieved from https://link.springer.com/book/10.1007%2F978-1-4020-4809-8

Alegre, M. & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40*, 41–61.

Altmann, E. G., Pierrehumbert, J. B. & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PLOS ONE, 6*(5), e19009.

Andringa, S. & Dabrowska, E. (2019). Individual differences in first and second language ultimate attainment and their causes. *Language Learning, 69*(S1), 5-12. doi:10.1111/lang.12328

Andringa, S. & Godfroid, A. (2019). Call for participation. *Language Learning, 69,* 5-10. doi:10.1111/lang.12338

Arnon, I. & Clark, E. V. (2011). Why brush your teeth is better than teeth: Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development, 7,* 107–129. doi:10.1080/15475441.2010.505489

Arnon, I. & Cohen Priva, U. (2013). More than words: the effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*(3), 349-371.

Arnon, I. & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language,* 62, 67-82.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M. & Zeschel, A. (2010). Cognitive corpus linguistics: five points of debate on current theory and methodology. *Corpora, 5*(1), 1–27.

Ashton, R. H. (2000). A review and analysis of research on the test–retest reliability of professional judgment. *Journal of Behavioral Decision Making, 13*(3), 277–294.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Baayen, R.H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5(3), 436-461*.

Baayen, R. H., Davidson, D. J. & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Baayen, R. H. & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*, 12-28.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P. & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438–482.

Baayen, R. H., Tomaschek, F., Gahl, S. & Ramscar, M. (2017). The Ecclesiastes principle in language change. In M. Hundt, S. Mollin & S. E. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (pp. 21–48). Cambridge: Cambridge University Press.

Backus, A. (2013). A usage-based approach to borrowability. In E. Zenner & G. Kristiansen (Eds.), *New perspectives on lexical borrowing* (pp. 19–39). Berlin: Mouton de Gruyter.

Backus, A. (2015). Rethinking Weinreich, Labov & Herzog from a usage-based perspective: Contact-induced change in Dutch Turkish. *Taal en tongval: Tijdschrift voor taalvariatie*, *67*(2), 275-306. https://doi.org/10.5117/TET2015.2.BACK

Backus, A. & Mos, M. (2011). Islands of (im)productivity in corpus data and acceptability judgments: Contrasting two potentiality constructions in Dutch. In: D. Schönefeld (Ed.), *Converging Evidence* (pp. 165-192). Amsterdam: John Benjamins.

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

Baker, F. B. & Seock-Ho, K. (2004). *Item response theory: Parameter estimation techniques*. New York: Dekker.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H. & Yap, M. J. (2004). Visual word recognition for single syllable words. *Journal of Experimental Psychology General, 133*(2), 283-316.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B. & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445-459.

Balota, D. A., Pilotti, M. & Cortesem, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition, 29*(4), 639–647.

Bannard, C. & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19,* 241–248.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*, 280–289. doi:10.1016/j.tics.2007.05.005

Bar, M., Neta, M. & Linz, H. (2006). Very first impressions. *Emotion,* 6(2), 269–278.

Bard, E., Robertson, D. & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72*, 32-68.

Barking, M., Backus, A. & Mos, M. (submitted). Comparing forward and reverse transfer from Dutch to German.

Barlow, M. (2013). Individual differences and usage-based grammar. *International Journal of Corpus Linguistics, 18*(4), 443–478.

Barlow, M. & Kemmer, S. (2000). *Usage-based models of language*. Cambridge: Cambridge University Press.

Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Barth, D. & Kapatsinski, V. (2014). A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of am, are and is. *Corpus Linguistics and Linguistic Theory, 13*(2), 203-260. doi:10.1515/cllt-2014-0022.

Bates, D. M., Mächler, M., Bolker, B. M. & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence*. New York: Springer.

Blasko, D. G. & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 295-308.

Blom, W. B. T., Paradis, J. & Sorenson Duncan, T. (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular -s in child L2 English. *Language Learning, 62*(3), 965-994.

Blom, E. & Unsworth, S. (2010). *Experimental methods in language acquisition research*. Amsterdam: Benjamins.

Boersma, P. & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.06, retrieved 21 February 2015 from http://www.praat.org/

Borensztajn, G., Zuidema, W. & Bod, R. (2009). Children's grammars grow more abstract with age -Evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science, 1*, 175-188. doi:10.1111/j.1756-8765.2008.01009.x

Bornkessel-Schlesewsky, I. & Schlesewsky, M. (2007). The wolf in sheep's clothing: Against a new judgment-driven imperialism. *Theoretical Linguistics, 33*(3), 319-333.

Branigan, H. P. & Pickering, M. J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences, 40*, 1-73. doi:10.1017/S0140525X16002028

Brothers, T., Swaab, T. Y. & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition*, *136*, 135–149.

Brothers, T., Swaab, T. Y. & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language, 93*, 203-216.

Bryman, A. (2004) Triangulation. In M. Lewis-Beck, A. Bryman & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods* (p. 1143). doi:10.4135/9781412950589.n1031

Buchstaller, I. & Khattab, G. (2013). Population samples. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 74-95). Cambridge: Cambridge University Press.

Burmester, J., Spalek, K. & Wartenburger, I. (2014). Context updating during sentence comprehension: The effect of aboutness topic. *Brain and Language, 137*, 62-76.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition, 24*, 215-221.

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language, 82*, 529–551.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, J. & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of 'don't' in English. *Linguistics, 37*, 575–596.

Cacioppo, J. T., Petty, R. E. & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307.

Caldwell-Harris, C., Berant, J. & Edelman, Sh. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In D. Divjak & S. Gries (Eds.), *Frequency effects in language representation* (pp. 165-194). Berlin: Mouton de Gruyter.

Camblin, C., Gordon, P. & Swaab, T. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language, 56*(1), 103-128.

Carlsson, K., Petrovic, P., Skare, S., Petersoon, K. M. & Ingvar, M. (2000). Tickling expectations: neural processing in anticipation of a sensory stimulus. *Journal of Cognitive Neuroscience, 12*(4), 691–703.

Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Boston: Houghton-Mifflin.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning, 33*(3), 343–377.

Chen, S. & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language, 13*, 359–394.

Chen, P. & Popovich, P. (2002). *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Christiansen, M. & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences 31*, 489-558.

Christiansen, M. & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39*, E62. doi:10.1017/S0140525X1500031X

Church, K. & Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering, 1*(2), 163–190.

Churchill, G. & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*(4), 360–375.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. doi:10.1017/S0140525X12000477

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

Colman, A. M., Norris, C. E. & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports, 80*(2), 355–362.

Connine, C. M., Mullennix, J., Shernoff, E. & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *16*, 1084-1096.

Crain, S. & Lillo-Martin, D. (1999). *An introduction to linguistic theory and language acquisition*. Malden: Blackwell.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. London: Longman.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronk, B. C., Lima, S. D. & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research, 22*(1), 59-82.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Dąbrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language*, *58*, 931-951.

Dąbrowska, E. (2010). Naive v. expert competence: An empirical study of speaker intuitions. *The Linguistic Review, 27*, 1–23.

Dąbrowska, E. (2010). The mean lean grammar machine meets the human mind: Empirical investigations of the mental status of rules. In H.-J. Schmid & S. Handl (Eds.), *Cognitive foundations of linguistic usage patterns. Empirical approaches* (pp. 151–170). Berlin: Mouton de Gruyter.

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism, 2*(3), 219-253.

Dąbrowska, E. (2013). Functional constraints, usage, and mental grammars: A study of speakers' intuitions about questions with long-distance dependencies. *Cognitive Linguistics, 24*(4), 633–665.

Dąbrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics, 25*(4), 617–653.

Dąbrowska, E. (2016a). Cognitive Linguistics' seven deadly sins. *Cognitive Linguistics, 27*(4), 479–491.

Dąbrowska, E. (2016b). Language in the mind and in the community. In J. Daems, E. Zenner, K. Heylen, D. Speelman & H. Cuyckens (Eds.), *Change of paradigms – New paradoxes* (pp. 221–235). Berlin: Walter de Gruyter.

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition, 178*, 222-235.

Dąbrowska, E. & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, *16*(3), 437-474.

Dambacher, M., Kliegl, R., Hofmann, M. & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103.

Daneman, M. E. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

De Deyne, S. & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods, 40*, 198-205.

De Bot, K. & Schrauf, R. (2009). *Language development over the lifespan*. New York: Routledge.

DeLong, K., Urbach, T. & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-1121.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology, 25*(2), 108–127.

Divjak, D. (2016). The role of lexical frequency in the acceptability of syntactic variation. Evidence from that-clauses in Polish. *Cognitive Science, 41*(2), 354-382. doi:10.1111/cogs.12335.

Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Eckert, P. (1997). Age as a sociolinguistic variable. In Florian Coulmas (Ed.), *Handbook of sociolinguistics* (pp. 151–167). Oxford: Blackwell.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology, 41*, 87–100.

Eckert, P. (2013). Ethics in linguistic research. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 11-26). Cambridge: Cambridge University Press.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188.

Ellis, N. C. & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics and education. *Corpus Linguistics and Linguistic Theory, 5*(1), 61-78.

Ellis, R. (1991). Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition, 13*(2), 161–186.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141–172.

Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical Linguistics, 33*, 269–318.

Fernandez Monsalve, I., Frank, S.L. & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398-408). Avignon, France: Association for Computational Linguistics.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M. & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from Chronolex. *Frontiers in Psychology, 2*(306), 1-10.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics: and sex and drugs and rock 'n' roll*. Los Angeles, CA: Sage.

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. London: Thousand Oaks.

Fitzpatrick, T., Playfoot, D., Wray, A. & Wright, M. (2015). Establishing the Reliability of Word Association Data for Investigating Individual and Group Differences. *Applied Linguistics, 36*, 23-50.

Flynn, S. (1986). Production vs. comprehension: Differences in underlying competences. *Studies in Second Language Acquisition, 8,* 135–164.

Forster, K. & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 627–635.

Foulkes, P. (2006). Phonological variation: A global perspective. In B. Aarts & A. McMahon (Eds.), *The Handbook of English Linguistics* (pp. 625–669). Oxford, UK: Blackwell.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science, 5,* 475-494.

Frank, S. L., Otten, L. J., Galli, G. & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain & Language*, *140*, 1–11.

Gardner, M. K., Rothkopf, E. Z., Lapan, R. & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition*, *15*, 24–28.

Garrod, S. & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8–11.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*, 256–281.

Gibbs, R. Jr. (2006). Introspection and cognitive linguistics: Should we trust our own intuitions? *Annual Review of Cognitive Linguistics, 4,* 135-151.

Gibson, E. & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences, 14*, 233–234.

Gibson, E. & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes, 28*(1), 88–124.

Gilquin, G. & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1), 1–26.

Goldberg, A. E. (2006). *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Goudbeek, M., Swingley, D. & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1913–1933.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford: Oxford University Press.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *international Journal of Corpus Linguistics, 13*(4), 403-437.

Gries, S. Th. (2010a). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: selected approaches* (pp. 269-291). Frankfurt am Main: Peter Lang.

Gries, S. Th. (2010b). Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon, 5*(3), 323-346.

Gries, S. Th. (2012). Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. *Studies in Language, 36*(3), 477-510.

Gries, S. Th. (2014). Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us. In I. Taavitsainen, M. Kytö, C. Claridge & J. Smith (Eds.), *Developments in English: expanding electronic evidence* (pp. 29–47). Cambridge: Cambridge University Press.

Gries, S. Th. (2015). Quantitative methods in linguistics. In J. D. Wright (Ed.), *International Encyclopedia of the Social and Behavioral Sciences,* 2nd edn., vol. 19 (pp. 725–732). Amsterdam: Elsevier.

Gries, S. T. & Divjak, D. (2012.). *Frequency effects in language representation*. Retrieved from https://ebookcentral.proquest.com

Gries, S. Th. & Newman, J. (2013). Creating and using corpora. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 257-287). Cambridge: Cambridge University Press.

Gries, S. Th. & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics, 7*, 163–186.

Griffin, Z.M. & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language, 38*, 313-338.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *56*, 1213-1236. doi:10.1080/02724980343000071

Gutiérrez-Clellen, V. & Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24*(2), 267-288.

Hammersley, M. (2008). Troubles with triangulation. In M. M. Bergman (Ed.), *Advances in mixed methods research* (pp. 22–36). London: SAGE.

Hashemi, M. R. & Babaii, E. (2013). Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal,* 97(4), 828–852.

Henrich, J., Heine, S. & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83. doi:10.1017/S0140525X0999152X

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428.

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Science, 6*(3), 253–271.

Howes, D. & Solomon, R. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*(6), 401-410.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626*, 118–135.

Hulstijn, J. (2010). Measuring second language proficiency. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 185-200). Amsterdam: Benjamins.

Isakov, A., Holcomb, A., Glowacki, L. & Christakis, N. A. (2016). Modeling the role of networks and individual differences in inter-group violence. *PloS ONE*, *11*(2), e0148314. https://doi.org/10.1371/journal.pone.0148314

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language,* 59(4), 434–446.

Janda, L. A. (2013). Quantitative methods in cognitive linguistics: An introduction. In L. A. Janda (Ed.), *Cognitive Linguistics: The quantitative turn* (pp. 1–32). Berlin: De Gruyter Mouton.

Janssen, N. & Barber, H.A. (2012). Phrase frequency effects in language production. *PLoS ONE, 7*(3): e33202. doi:10.1371/journal.pone.0033202

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87-106.

Jiang, X. L. & Cillessen, A. (2005). Stability of continuous measures of sociometric status: A meta-analysis. *Developmental Review,* 25(1), 1–25.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R $^{2}{}_{\text{GLMM}}$ to random slopes models. *Methods in Ecology and Evolution, 5*, 944–946. doi:10.1111/2041-210X.12225

Johnson, J. S., Shenkman, K. D., Newport, E. L. & Medin, D. L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language, 35*(3), 335–352.

Jolsvai, H., McCauley, S. M. & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multiword chunks. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*. Retrieved from https://escholarship.org/uc/item/5cv7b5xs

Juhasz, B. J., Lai, Y.-H. & Woodcock, M. L. (2015). A database of 629 English compound words: Ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods, 47*(4), 1004-1019.

Juhasz, B. J. & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1312–1318.

Jurafsky, D., Bell, A., Gregory, M. & Raymond, W. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254_. Amsterdam: John Benjamins.

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In R. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 135-168). Cambridge: Cambridge University Press.

Kamoen, N. (2012). *Positive versus negative: A cognitive perspective on wording effects for contrastive questions in attitude surveys*. Utrecht: LOT dissertation.

Keller, F. & Alexopoulou, T. (2001). Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition, 79*, 301–372.

Kemp, N., Mitchell, P. & Bryant, P. (2017). Simple morphological spelling rules are not always used: Individual differences in children and adults. *Applied Psycholinguistics, 38*, 1071-1094. doi:10.1017/S0142716417000042

Kertész, A., Schwarz-Friesel, M. & Consten, M. (2012). Introduction: converging data sources in cognitive linguistics. *Language Sciences, 34*(6), 651–655.

Keuleers, E. & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: an overview of recent developments introduction. *Quarterly Journal of Experimental Psychology*, *68*(8), 1457-1468.

Keuleers, E., Diependaele, K. & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1,* 174-189. http://doi.org/10.3389/fpsyg.2010.00174

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*(1), 1-37.

Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283–312). San Diego, CA: Academic Press.

Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284.

Koole, S. L. & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*, 608–614. doi:10.1177/1745691612462586

Kristiansen, G. & Dirven, R. (2008). *Cognitive sociolinguistics: language variation, cultural models, social systems.* Berlin: Mouton de Gruyter.

Kuhl, P. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America, 97*(22), 11850–11857.

Kuperberg, G. R. & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32-59. doi:10.1080/23273798.2015.1102299

Kutas, M., DeLong, K. A. & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the Brain: Using Our Past to Generate a Future* (pp. 190–207). New York: Oxford University Press.

Labov, W. (n.d.). Some observations on the foundations of linguistics. Retrieved from http://www.ling.upenn.edu/~wlabov/Papers/Foundations.html (11 July, 2012).

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

Labov, W. (2001). *Principles of linguistic change. Vol. 2: Social factors*. Oxford: Blackwell.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: Chicago University Press.

Langacker, R. (1987). *Foundations of Cognitive Grammar, Vol. I.* Stanford: Stanford University Press.

Langacker, R. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp. 1-63). Stanford, CA: CSLI Publications.

Langsford, S., Perfors, A., Hendrickson, A., Kennedy, L. & Navarro, D. (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics, 3*(1), 1–34.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177. doi:10.1016/j.cognition.2007.05.006

Levon, N. (2013). Surveys and interviews. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 96-115). Cambridge: Cambridge University Press.

Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children’s production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*(3), 481-507.

Lilly, B. (2009). Optimizing stimuli order in marketing experiments: A comparison of four orders using six criteria. *Journal of Targeting, Measurement and Analysis for Marketing, 17*, 245-255. doi:10.1057/jt.2009.17

Linzen, T. & Jaeger, F. (2016). Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. *Cognitive Science, 40*(6), 1382–1411.

Lüdeling, A. & Kytö, M. (2008). *Corpus Linguistics: An International Handbook. Volume 1.* Berlin: Mouton de Gruyter.

Luka, B. & Barsalou, L. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language, 52*(3), 436-459.

Mandera, P., Keuleers, E. & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language*, *94*, 305-315.

Maxwell, S. E., Kelley, K. & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–63.

McCauley, S. & Christiansen, M. (2014). Acquiring formulaic language: A computational model. *The Mental Lexicon, 9*, 419-436.

McCauley, S., Isbilen, E. & Christiansen, M. (2017). Chunking ability shapes sentence processing at multiple levels of abstraction. In G. Gunzelmann, A. Howes, T. Tenbrink & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2681– 2686). Austin, TX: Cognitive Science Society.

McDonald, S. A. & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science, 14*(6), 648-652.

McEvoy, C. L. & Nelson, D. L. (1982). Category name and instance norms for 106 categories of various sizes. *American Journal of Psychology*, *95*, 581-634.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Hove, England: Psychology Press.

McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547-559.

Meng, M. & Bader, M. (2000). Ungrammaticality detection and garden-path strength: Evidence for serial parsing. *Language and Cognitive Processes, 15*, 615–666.

Misyak, J. B. & Christiansen, M. H. (2012). Statistical learning and language: an individual differences study. *Language Learning*, *62*(1), 302–331. doi:10.1111/j.1467-9922.2010.00626.x

Misyak, J. B., Christiansen, M. H. & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science, 2*, 138–153. doi:10.1111/j.1756-8765.2009.01072.x

Mos, M., van den Bosch, A. & Berck, P. (2012). The predictive value of word-level perplexity in human sentence processing: A case study on fixed adjective-preposition constructions in Dutch. In S. Th. Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing* (pp. 207–239). Berlijn: De Gruyter.

Myers, J., Well, A. & & Lorch, R. (2010). *Research design and statistical analysis*. New York: Routledge.

Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/

Nordquist, D. (2009). Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory, 5*(1), 105–130.

Nosek, B. & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137-141.

Nurius, P. S. & Macy, R. J. (2008). Heterogeneity among violence-exposed women: Applying person-oriented research methods. *Journal of Interpersonal Violence*, *23*(3), 389-415.

Olsthoorn, N., Andringa, S. & Hulstijn, J. (2014). Visual and auditory digit-span performance in native and nonnative speakers. *International Journal of Bilingualism*, *18*(6), 663-673.

Oostdijk, N., Reynaert, M., Hoste, V. & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (eds.), *Essential speech and language technology for Dutch: Theory and applications of natural language processing* (pp. 219–247). Dordrecht: Springer.

Otten, M. & Van Berkum, J. (2008). Discourse-based anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*, 464–496.

Paiva, C., Barroso, E., Carneseca, E., de Pádua Souza, C., Thomé dos Santos, F., López, R. & Sakamoto Ribeiro Paiva, B. (2014). A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. *BMC Medical Research Methodology, 14*(1), 8–18.

Paradis, J. (2010). Comparing typically-developing children and children with specific language impairment. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 223-244). Amsterdam: Benjamins.

Pearl, L. (2010). Using computational modeling in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 163-184). Amsterdam: Benjamins.

Pickering, M. J. & Ferreira, V. S. (2008). Structural priming: A critical review. *Psycholinguistic Bulletin, 134*(3), 427–459.

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

Podesva, R. J. & Sharma, D. (2013). *Research methods in linguistics*. Cambridge: Cambridge University Press.

Popiel, S. J. & McRae, K. (1988). The figurative and literal senses of idioms; or, all idioms are not used equally. *Journal of Psycholinguistic Research, 17*, 475–487.

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15.

Princeton University (2010). About WordNet. Retrieved from https://wordnet.princeton.edu/

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Rastle, K., Harrington, J. & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology, 55*(4), 1339-1362.

Raven, J., Raven, J. C. & Court, J. H. (1998). *Raven manual section 4: Advanced progressive matrices*. Oxford, UK: Oxford Psychologists Press.

Rayner, K., Ashby, J., Pollatsek, A. & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z reader model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720–732.

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, held in conjunction with The 38th Annual Meeting of the Association for Computational Linguistics*, 1-6.

Roark, B., Bachrach, A., Cardenas, C. & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore), 324–333.

Roehr, K. (2008). Linguistic and metalinguistic categories in second language learning. *Cognitive Linguistics, 19*, 67–106.

Roland, D., Yun, H., Koenig, J.-P. & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, *122*, 267–279.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110-114.

Sampson, G. R. (2007). Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory, 3*(1), 1-32.

Sankoff, G. (2006). Age: Apparent time and real time. In K. Brown (Ed.), *The encyclopedia of language and linguistics*, 2nd edn., vol. 1 (pp. 110–116). Oxford: Elsevier.

Schäfer, R. & Bildhauer. F. (2012). Building large corpora from the web using a new efficient tool chain. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12) (pp. 486–493). Istanbul, Turkey: European Language Resources Association.

Schilling, N. (2013). Surveys and interviews. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 96-115). Cambridge: Cambridge University Press.

Schmid, H.-J. (2007). Entrenchment, salience and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 117-138). Oxford: Oxford University Press.

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 101-133). Berlin: Mouton de Gruyter.

Schmid, H.-J. (2015). A blueprint of the Entrenchment-and-Conventionalization Model. Y*earbook of the German Cognitive Linguistics Association*, *3*, 1-27.

Schmid, H-J. & Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics, 24*(3), 531-577.

Schmid, H.-J. & Mantlik, A. (2015). Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles. *Anglia, 133*(4), 583–623. doi:10.1515/ang-2015-0056

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90–100. doi:10.1037/a0015108

Schönefeld, D. (2011). *Converging evidence: Methodological and theoretical issues for linguistic research.* Amsterdam: John Benjamins.

Schönefeld, D. (2011). Introduction: On evidence and the convergence of evidence in linguistic research. In D. Schönefeld (Ed.), *Converging evidence. Methodological and theoretical issues for linguistic research* (pp. 1–32). Amsterdam: John Benjamins

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

Schütze, C. T. & Sprouse, J. (2013). Judgment data. In R.J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 27-50). Cambridge: Cambridge University Press.

Sebregts, K. (2015). *The sociophonetics and phonology of Dutch r*. Utrecht: Utrecht University dissertation.

Seidenberg, M. S. (1997). Language acquisition and use: learning and applying probabilistic constraints. *Science, 275*, 1599–1603.

Seto, E., Hua, J., Wu, L., Shia, V., Eom, S., Wang, M. & Li, Y. (2016). Models of individual dietary behavior based on smartphone data: the influence of routine, physical activity, emotion, and food environment. *PloS ONE*, *11*(4), e0153085. https://doi.org/10.1371/journal.pone.0153085

Shaoul, C., Baayen, R. H., and Westbury, C. F. (2014). N-gram probability effects in a cloze task. *The Mental Lexicon, 9*, 437-472.

Shaoul, C., Westbury, C. F. & Baayen, R. H. (2013). The subjective frequency of word n-grams. *Psihologija, 46*(4), 497–537.

Sharma, D. (2011). Style repertoire and social change in British Asian English. *Journal of Sociolinguistics, 15*(4), 464–492.

Simmons, W. K., Martin, A. & Barsalou, L. W. (2005) Pictures of appetizing foods activate gustatory cortices for taste and reward. *Cerebral Cortex, 15*, 1602–1608.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia A., Conklin, K., van Heuven, W. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 776-784.

Smith, N. J. & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1637–1642). Austin, TX: Cognitive Science Society.

Smith, N. J. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319. doi:10.1016/j.cognition.2013.02.013

Smits, R., Sereno, J. & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 733–754.

Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. *Language, 76*, 859–890.

Sprouse, J. (2008). Magnitude estimation and the non-linearity of acceptability judgments. In N. Abner & J. Bishop (Eds.), *West Coast Conference on Formal Linguistics* (WCCFL) (pp. 397–403). Somerville, MA: Cascadilla Proceedings Project.

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

Sprouse, J. & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics, 48*(3), 609–652.

Sprouse, J., Schütze, C. T. & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua, 134*, 219-248.

Stanovich, K. E. & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*, 402-433.

Stefanowitsch, A. & Gries, S.Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*, 209-243.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. *Proceedings of the International Conference on Spoken Language Processing* (pp. 901–904). Denver, Colorado.

Street, J., & Dąbrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua, 120*(8), 2080-2094.

Street, J., & Dabrowska, E. (2014). Lexically specific knowledge and individual differences in adult native speakers' processing of the English passive. *Applied Psycholinguistics*, *35*(1), 97-118.

Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*, 209-243.

Stubbs, M. (1993). British traditions in text analysis: From Firth to Sinclair. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In honour of John Sinclair* (pp. 1-33). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Schwanenflugel, P. J. & Gaviska, D. C. (2005). Psycholinguistic aspects of word meaning. In D. A. Cruse, F. Hundsnurscher, M. Job & P. R. Lutzeier (Eds.), *Lexikologie: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wordschätzen* [Lexicology: An international handbook on the nature and structure of words and vocabularies] (pp. 1735-1748). Berlin: Mouton de Gruyter.

Tabatabaei, O. & Dehghani, M. (2012). Assessing the reliability of grammaticality judgment tests. *Procedia – Social and Behavioral Sciences, 31*, 173–182.

Tabossi, P., Fanari, R. & Wolf, K. (2009). Why are idioms recognized fast? *Memory & Cognition, 37*(4), 529–540.

Tanner, D., Inoue, K. & Osterhout, L. (2014). Brain-based individual differences in on-line L2 grammatical comprehension. *Bilingualism, Language and Cognition, 17*(2), 277-293.

Taylor, J. R. (2012). *The Mental Corpus. How language is represented in the mind*. Oxford: Oxford University Press.

Taylor, W. L. (1953). 'Cloze' procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415-433.

Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgement tasks. *Cognitive Development, 19*(1), 15–34.

Tily, H., Gahl, S., Arnon, I., Kothari, A., Snider, N. & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language & Cognition, 1*, 147-165.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Traxler, M. J. & Foss, D. J. (2000). Effects of sentence constraint on priming in natural language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1266-1282. doi:10.1037/0278-7393.26.5.1266

Treadwell, C. (2017). *Introducing communication research. Paths of inquiry* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Tremblay, A. & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language; Acquisition and communication* (pp. 151–173). London: The Continuum International Publishing Group.

Tremblay, A. & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon, 6*(2), 302–324.

Tryk, H. (1968). Subjective scaling of word frequency. *American Journal of Psychology*, *81*(2), 170-177.

Tvesis, C. (2008). *Hope over fear* [Mosaic Illustration]. Retrieved from http://www.dripbook.com/tsevis/illustration-portfolio/barack-obamai/#288337 (6 February, 2014).

University of Twente, Human Media Interaction (n.d.). Twente News Corpus (TwNC): A multifaceted Dutch news corpus. Retrieved from http://hmi.ewi.utwente.nl/TwNC/description

Unsworth, S. & Blom, E. (2010). Comparing L1 children, L2 children and L2 adults. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 201-222). Amsterdam: Benjamins.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V. & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443–467.

Van den Bemd, E., Mos, M., Alishahi, A. & Shayan, S. (2014). Does sentence structure boost early word learning? An artifical language learning study. *Wiener Linguistische Gazette*, *78*(A), 103-119.

VanGeest, J., Wynia, M., Cummins, D. & Wilson, I. (2002). Measuring deception: test-retest reliability of physicians' self-reported manipulation of reimbursement rules for patients. *Medical Care Research and Review, 59*(2), 184–196.

Verhagen, A. (2003). Hoe het Nederlands zich een eigen weg baant: Vergelijkende en historische observaties vanuit een constructie-perspectief. *Nederlandse Taalkunde, 8*, 328-346.

Verhagen, A. (2005). Constructiegrammatica en 'usage based' taalkunde. *Nederlandse Taalkunde, 10*, 197-222.

Verhagen, V. & Backus, A. (2011). Individual differences in the perception of entrenchment of multiword units: Evidence from a Magnitude Estimation task. Toeg*epaste Taalwetenschap in Artikelen [Applied Linguistics in Article Form], 84/85,* 155–165.

Vindras, P., Desmurget, M. & Baraduc, P. (2012). When one size does not fit all: a simple statistical method to deal with across-individual variations of effects. *PLoS ONE*, *7*(6), e39059. https://doi.org/10.1371/journal.pone.0039059

von Eye, A. & Bogat, G. A. (2006). Person-oriented and variable-oriented research: Concepts, results, and development. *Merrill-Palmer Quarterly, 52*(3), 390-420.

von Eye, A., Bogat, G. A. & Rhodes, J. E. (2006). Variable-oriented and person-oriented perspectives of analysis: The example of alcohol consumption in adolescence. *Journal of Adolescence*, *29*(6), 981-1004.

von Eye, A., Mun, E. Y. & Indurkhya, A. (2004). Typifying developmental trajectories: A decision-making perspective. *Psychology Science, 46,* 65–98.

Wasow, T. & Arnold, J. (2005). Intuitions in linguistic argumentation. *Lingua, 115*, 1481–1496.

Waters, G. & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Quarterly Journal of Experimental Psychology, 49*, 51-79.

Wechsler, D. (1981). *The Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J. & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*, 250–271. doi:10.1016/j.cogpsych.2008.08.002

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956–972.

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

Westfall, J., Kenny, D. & Judd, C. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020-2045.

Wiechmann, D. (2008). On the computation of collostruction strength. *Corpus Linguistics and Linguistic Theory, 4*(2), 253-290.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. & Bosch, A. van den. (2016). Prediction during natural language comprehension. *Cerebral Cortex, 26*, 2506-2516. doi:10.1093/cercor/bhv075

Williams, R. & Morris, R. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16*, 312–339.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, England: Cambridge University Press.

Wulff, S. (2009). Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory, 5*(1), 131–159.

Zachary, R. (1994). *Shipley Institute of Living Scale, revised manual*. Los Angeles: Weston Psychological Services.

Zimmerer. V., Cowell, P., Varley, R. (2011). Individual behavior in learning of an artificial grammar. *Memory and Cognition, 39*, 491–501. doi:10.3758/s13421-010-0039-y

Zipf, G. K. (1935). *The psychobiology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin Company.

Appendices

Appendix 2.1 Stimuli

	– context	+ context
1	naar huis home	Hij is gisteren vroeg naar huis gegaan. He went home early yesterday.
2	op school at school	Met die jongen heb ik vroeger op school gezeten. I went to school with that boy.
3	op vakantie on vacation	De buren zijn vorige week op vakantie gegaan. The neighbors went on vacation last week.
4	in de klas in the classroom	De jongen zit in de klas naast zijn beste vriend. In the classroom the boy sits next to his best friend.
5	in de tuin in the garden	De man is in de tuin aan het werk. The man is working in the garden.
6	in de keuken in the kitchen	Zijn moeder was in de keuken bezig met het avondeten. His mother was busy with the evening meal in the kitchen.
7	in de auto in the car	We hebben de boodschappen in de auto gelegd. We put the shopping in the car.
8	in bed in bed	Zij is nog niet in bed gaan liggen. She has not yet lain down in bed.
9	in de kamer in the room	Hij heeft het nieuwe schilderij in de kamer opgehangen. He has hung the new painting in the room.
10	aan tafel at table	De jongen zit aan tafel zijn ontbijt te eten. The boy is seated at the table eating his breakfast.
11	op de bank on the couch	De jongens liggen op de bank televisie te kijken. The boys are lying on the couch watching tv.
12	in slaap asleep	Ik kon vannacht niet in slaap vallen. I couldn’t fall asleep last night.

13	in het water in the water	De kinderen zijn in het water aan het spelen. The children are playing in the water.
14	in de lucht in the air	De maatregelen hebben al langer in de lucht gehangen. The measures have been up in the air for a while.
15	in de hand in the hand	Ze hebben zelf in de hand wat er gaat gebeuren. It's in their own hands what will happen.
16	in de winkel in the shop	Zijn nieuwe CD is in de winkel te koop. His new CD is for sale in the shops.
17	in de kerk in the church	Mijn ouders zijn in de kerk getrouwd. My parents got married in church.
18	in de bus in the bus	Het meisje zit in de bus met haar moeder. The girl is sitting in the bus with her mother.
19	aan de beurt be next	Hij wacht totdat hij aan de beurt is. He waits until it is his turn.
20	op de televisie on the television / on tv	Ze keken een film die op de televisie werd uitgezonden. They watched a movie that was broadcast on tv.
21	op de foto in the picture	De fotograaf zorgde ervoor dat iedereen op de foto stond. The photographer made sure everybody was in the picture.
22	naar de wc to the loo	De helft van de klas ging naar de wc in de pauze. Half the class went to the loo in the interval.
23	in het bos in the forest	Er wonen in het bos veel dieren. There are a lot of animals living in the forest.
24	op de hoek at the corner	De winkel bevindt zich op de hoek van de straat. The shop is located at the corner of the street.
25	in de kast in the cupboard	Ze heeft de spulletjes in de kast gelegd. She put the things in the cupboard.
26	in de oven in the oven	Ze staat op het punt de appeltaart in de oven te zetten. She's about to put the apple pie in the oven.

27	in bad in (the) bath	Het kindje werd door zijn moeder in bad gezet. The little child was put in (the) bath by his mother.
28	op de deur on the door	Er wordt op de deur geklopt. There's a knock at the door.
29	achter de computer behind the computer	De jongens zitten veel achter de computer volgens hun moeder. The boys spend a lot of time at the computer. according to their mother.
30	in de film in the film	De actrice gaat de hoofdrol in de film vertolken. The actress will play the leading part in the film.
31	in het licht in the light	Zij waarschuwde hem niet recht in het licht te kijken. She warned him not to look straight into the light.
32	in de pan in the pan	Je moet de groenten in de pan doen en even laten koken. You have to put the vegetables in the pan and let them boil for a while.
33	op de muur on the wall	Het filmpje werd op de muur geprojecteerd. The film was projected on the wall.
34	in de kring in the ring	De kinderen praten in de kring over het weekend. The children talked about the weekend in the ring.
35	van het dak off the roof / of the roof	De tuinman bood aan de bladeren van het dak te vegen. The gardener offered to sweep the leaves off the roof.
36	in het bed in the bed	De jongen lag nog in het bed toen zijn moeder binnenkwam. The boy was still lying in the bed when his mother came in.
37	tegen mama to mom	Hij zei tegen mama dat hij niet ging. He told mom he didn't go.

38	in de buik in the stomach	Hij vertelde de dokter dat hij pijn in de buik heeft. He told the doctor he has pain in the stomach.
39	met de hond with the dog	Mijn ouders gaan met de hond wandelen. My parents are going to take the dog for a walk.
40	in de bak in the bin / in jail	Criminelen horen in de bak te zitten. Criminals should be in jail.
41	in het paleis in the palace	De bruiloft werd in het paleis gevierd. The wedding was celebrated in the palace.
42	in het hoofd in the head	Hij had een plan in het hoofd toen hij vertrok. He had a plan in mind when he left.
43	in het bad in the bath	Het water in het bad bubbelt. The water in the bath is bubbling.
44	in zijn eentje on his own	Hij gaat in zijn eentje zitten. He is going to sit by himself.

Appendix 2.2 Raw frequencies and base-10 logarithms of the frequency of occurrence per million words in the Corpus of Spoken Dutch (CGN) for the noun (lemma search) and the specific phrase as a whole; mean familiarity ratings and standard deviations both at Time 1 and Time 2.

						Time 1				Time 2			
		Freq	Log	Freq	Log	−context		+context		−context		+context	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
1	naar huis	4730	2.70	1066	2.05	0.75	(.54)	0.71	(.64)	0.74	(.64)	0.69	(.67)
2	op school	3572	2.58	742	1.89	0.65	(.72)	0.68	(.66)	0.68	(.67)	0.68	(.52)
3	op vakantie	1715	2.26	480	1.70	0.85	(.51)	0.73	(.62)	0.92	(.56)	0.80	(.58)
4	in de klas	1484	2.19	341	1.56	0.33	(.61)	0.10	(.69)	0.39	(.64)	0.22	(.64)
5	in de tuin	873	1.96	251	1.42	0.29	(.68)	0.31	(.63)	0.30	(.69)	0.26	(.64)
6	in de keuken	727	1.88	223	1.37	0.40	(.69)	0.53	(.50)	0.36	(.66)	0.51	(.54)
7	in de auto	2811	2.47	207	1.34	0.40	(.64)	0.54	(.58)	0.41	(.74)	0.29	(.62)
8	in bed	1290	2.13	194	1.31	0.69	(.70)	0.28	(.88)	0.67	(.83)	0.56	(.65)
9	in de kamer	1941	2.31	190	1.30	0.27	(.68)	0.12	(.78)	0.17	(.85)	0.01	(.73)
10	aan tafel	1233	2.11	187	1.30	0.58	(.61)	0.48	(.68)	0.53	(.62)	0.35	(.66)
11	op de bank	877	1.97	172	1.26	0.61	(.49)	0.53	(.68)	0.54	(.70)	0.57	(.57)
12	in slaap	341	1.56	167	1.25	-0.18	(1.08)	0.28	(.90)	0.24	(.98)	0.46	(.76)
13	in het water	1959	2.31	148	1.20	0.18	(.70)	0.10	(.75)	0.11	(.77)	0.09	(.81)

						Time 1				Time 2			
		Freq	Log	Freq	Log	−context		+context		−context		+context	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
14	in de lucht	636	1.83	147	1.19	-0.05	(.78)	-0.92	(1.07)	-0.25	(.89)	-0.34	(.91)
15	in de hand	3062	2.51	141	1.17	-0.59	(.97)	-0.04	(.88)	-0.54	(1.00)	-0.19	(.94)
16	in de winkel	838	1.95	136	1.16	0.45	(.55)	0.30	(.70)	0.31	(.66)	0.25	(.79)
17	in de kerk	961	2.01	119	1.10	-0.35	(.93)	0.04	(.88)	-0.17	(.96)	-0.05	(.87)
18	in de bus	995	2.02	118	1.10	0.30	(.68)	0.46	(.63)	0.30	(.74)	0.31	(.62)
19	aan de beurt	294	1.49	108	1.06	0.16	(.69)	0.38	(.55)	0.18	(.73)	0.36	(1.18)
20	op de televisie	617	1.81	69	0.87	-0.27	(1.09)	-0.18	(1.09)	-0.20	(1.10)	-0.42	(1.29)
21	op de foto	1439	2.18	69	0.87	0.43	(.63)	0.46	(.63)	0.39	(.69)	0.44	(.67)
22	naar de wc	264	1.45	67	0.85	0.58	(.60)	0.56	(.62)	0.60	(.65)	0.38	(.70)
23	in het bos	493	1.72	66	0.85	0.17	(.74)	0.19	(.73)	0.13	(.74)	0.00	(.91)
24	op de hoek	606	1.81	65	0.84	-0.21	(.85)	0.13	(.77)	-0.27	(.89)	0.12	(.74)
25	in de kast	600	1.80	62	0.82	0.10	(.72)	0.35	(.66)	0.06	(.78)	0.24	(.59)
26	in de oven	223	1.37	60	0.81	0.23	(1.02)	0.36	(.55)	0.25	(.70)	0.42	(.62)
27	in bad	181	1.28	54	0.76	0.29	(.75)	0.06	(.81)	0.41	(.72)	0.15	(.72)
28	op de deur	1495	2.20	51	0.74	-0.64	(.97)	0.20	(.80)	-0.39	(.96)	0.02	(.82)
29	achter de	1099	2.06	41	0.65	0.47	(.73)	0.34	(.73)	0.50	(.68)	0.39	(.74)

						Time 1				Time 2			
		Freq	Log	Freq	Log	−context		+context		−context		+context	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
30	in de film	1658	2.24	33	0.55	-0.50	(1.06)	0.27	(.70)	-0.47	(.99)	0.17	(.80)
31	in het licht	1340	2.15	32	0.54	-0.53	(.85)	-0.18	(.74)	-0.46	(.86)	-0.24	(.78)
32	in de pan	214	1.35	29	0.50	0.05	(.74)	0.45	(.53)	0.15	(.66)	0.23	(.65)
33	op de muur	782	1.92	28	0.48	-0.50	(.99)	0.08	(.67)	-0.35	(.83)	-0.02	(.64)
34	in de kring	228	1.38	27	0.47	-0.61	(.88)	-0.30	(.98)	-0.46	(.87)	-0.41	(1.17)
35	van het dak	423	1.65	24	0.42	-1.05	(.97)	-0.33	(.87)	-0.68	(.96)	-0.24	(.80)
36	in het bed	1290	2.13	23	0.40	-0.51	(1.24)	-1.16	(1.29)	-0.51	(1.23)	-1.11	(1.43)
37	tegen mama	1188	2.10	22	0.38	-0.61	(1.10)	-0.24	(1.13)	-0.48	(1.18)	-0.37	(1.13)
38	in de buik	286	1.48	18	0.30	-1.24	(.96)	-1.65	(1.15)	-1.37	(1.03)	-1.50	(1.13)
39	met de hond	789	1.92	15	0.23	-0.19	(1.06)	0.27	(1.23)	-0.30	(.93)	0.06	(.89)
40	in de bak	331	1.54	15	0.23	-0.91	(1.02)	-0.38	(.91)	-0.82	(1.06)	-0.31	(.99)
41	in het paleis	160	1.23	13	0.17	-0.71	(1.02)	-0.37	(.91)	-0.77	(1.01)	-0.42	(1.07)
42	in het hoofd	1700	2.25	13	0.17	-1.27	(1.09)	-1.57	(1.03)	-1.14	(1.17)	-1.67	(1.17)
43	in het bad	181	1.28	12	0.14	-0.64	(1.11)	-0.40	(.91)	-0.64	(1.00)	-0.44	(.99)
44	in zijn eentje	29	0.50	9	0.02	-0.28	(.95)	0.01	(.88)	-0.27	(.90)	-0.11	(.96)

Appendix 3.1 Stimuli in the order of presentation

1	naar huis	home
2	uit de kast	from the cupboard; out of the closet
3	bij de fietsen	near the bicycles
4	op papier	on paper
5	in de groente	in the vegetables
6	onder de wol	underneath the wool; turn in
7	op het boek	on the book; on top of the book
8	onder de mat	underneath the mat
9	onder het asfalt	underneath the asphalt
10	in de shampoo	in the shampoo
11	in het geld	in the money (*zwemmen in het geld* 'have pots of money')
12	langs de auto	past the car
13	in het algemeen	in general
14	op vakantie	on vacation
15	in de winkel	in the shop
16	in het bos	in the forest
17	op de bon	on the ticket (also: be booked; rationed)
18	naast het hek	beside the fence
19	voor de schommel	in front of the swing
20	langs de boeken	along the books
21	in de lucht	in the air
22	tot morgen	till tomorrow
23	in de klas	in the classroom
24	in de pan	in the pan
25	in de kamer	in the room
26	uit de kom	from the bowl; out of its socket
27	in de oven	in the oven
28	in de bak	in the bin; in jail
29	in de piano	in the piano
30	naast de bloemen	beside the flowers
31	voor de juf	for the teacher/Miss
32	naast het café	beside the cafe
33	tegen de vlakte	against the plain (*tegen de vlakte gaan* 'be knocked down')

34	uit de gang	from the corridor
35	naar de boom	towards the tree
36	op de pof	on tick
37	tegen de grond	against the ground; to the ground
38	onder de dekens	underneath the blankets
39	over de kop	over the head (*over de kop gaan* 'overturn' and 'go broke'; *zich over de kop werken* 'work oneself to death')
40	rond de middag	around midday
41	onder elkaar	amongst themselves; by ourselves; one below the other
42	van het dak	off the roof; of the roof
43	aan tafel	at table
44	naar de wc	to the loo
45	langs het park	along the park
46	met gemak	with ease
47	op televisie	on the television; on tv
48	naast de auto	beside the car
49	in het donker	in the dark
50	om de tekeningen	for the drawings; around the drawings
51	in de tuin	in the garden
52	in de oren	in the ears (*iets in de oren knopen* 'get something into one's head; *gaatjes in de oren hebben* 'have pierced ears')
53	langs het water	along the water
54	in bad	in (the) bath
55	in de koffie	in the coffee
56	tegen mama	to mom; against mom
57	over de streep	across the line (*iemand over de streep trekken* 'win someone over')
58	in het paleis	in the palace
59	uit de kunst	out of the art; amazing
60	in de bus	in the bus
61	op de bank	on the couch
62	op de hoek	at the corner

63	met het doel	with the goal (*met het doel om* 'with a view to')
64	over het gras	across the grass; about the grass
65	over het karton	over the cardboard; about the cardboard
66	in de keuken	in the kitchen
67	met de schoen	with the shoe
68	op de film	on (the) film
69	op de meester	on the teacher/master; at the teacher/master
70	in de kast	in the cupboard
71	aan de beurt	be next
72	langs de tafel	along the table
73	uit het niets	out of nothingness
74	in de auto	in the car
75	in de rondte	in a circle
76	in de foto	in the picture
77	op school	at school
78	rond de ingang	around the entrance
79	uit de trommel	from the tin box; out of the tin box

Appendix 3.2 Raw frequency and base-10 logarithms of the frequency of occurrence per million words in the subset of the corpus SoNaR* for the noun (lemma search) and the specific phrase as a whole; mean familiarity ratings and standard deviations both at Time 1 and Time 2.
* This subset consists of texts originating from the Netherlands (143.8 million words) and texts originating either from the Netherlands or Belgium (51.8 million words).

						Time 1				Time 2			
		Freq	Log	Freq	Log	Likert		ME		Likert		ME	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
1	naar huis	84918	2.64	14688	1.88	0.94	(0.41)	1.17	(0.50)	1.02	(0.50)	1.36	(0.51)
77	op school	58222	2.47	8543	1.64	0.81	(0.34)	1.15	(0.79)	0.95	(0.37)	1.00	(0.61)
13	in het algemeen	37893	2.29	5778	1.47	0.54	(0.74)	0.97	(0.42)	0.65	(0.65)	0.87	(0.64)
21	in de lucht	17713	1.96	4485	1.36	0.61	(0.38)	0.47	(0.48)	0.65	(0.43)	0.63	(0.43)
61	op de bank	28615	2.17	4221	1.33	0.86	(0.36)	0.93	(0.56)	0.89	(0.66)	1.06	(0.57)
14	op vakantie	15864	1.91	3742	1.28	0.86	(0.39)	1.14	(0.43)	0.88	(0.46)	1.07	(0.45)
74	in de auto	37927	2.29	3532	1.26	0.84	(0.36)	0.90	(0.44)	0.77	(0.53)	0.88	(0.62)
25	in de kamer	44194	2.35	3259	1.22	0.59	(0.46)	0.94	(0.49)	0.77	(0.41)	0.74	(0.62)
51	in de tuin	10213	1.72	2860	1.17	0.65	(0.58)	0.60	(0.60)	0.64	(0.57)	0.83	(0.55)
4	op papier	11249	1.76	2606	1.12	0.80	(0.30)	0.82	(0.53)	0.75	(0.42)	0.75	(0.61)

						Time 1				Time 2			
		Freq	Log	Freq	Log	Likert		ME		Likert		ME	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
43	aan tafel	2827	1.16	2439	1.10	0.76	(0.32)	0.89	(0.44)	0.79	(0.58)	0.98	(0.51)
66	in de keuken	7584	1.59	2174	1.05	0.72	(0.47)	0.92	(0.40)	0.68	(0.58)	0.92	(0.48)
47	op televisie	12003	1.79	1955	1.00	0.82	(0.49)	1.02	(0.60)	0.83	(0.55)	1.18	(0.47)
23	in de klas	7181	1.56	1924	0.99	0.69	(0.43)	0.93	(0.57)	0.69	(0.63)	0.80	(0.65)
22	tot morgen	46260	2.37	1820	0.97	0.91	(0.36)	1.36	(0.55)	1.08	(0.50)	1.32	(0.44)
71	aan de beurt	7759	1.60	1743	0.95	0.71	(0.42)	0.55	(0.64)	0.70	(0.49)	0.77	(0.51)
15	in de winkel	12870	1.82	1611	0.92	0.74	(0.50)	1.05	(0.45)	0.82	(0.52)	0.82	(0.61)
60	in de bus	12053	1.79	1533	0.89	0.71	(0.33)	0.68	(0.59)	0.66	(0.48)	0.60	(0.73)
49	in het donker	13022	1.82	1521	0.89	0.78	(0.27)	0.76	(0.43)	0.75	(0.45)	0.90	(0.37)
16	in het bos	16681	1.93	1295	0.82	0.53	(0.49)	0.83	(0.55)	0.57	(0.59)	0.61	(0.58)
54	in bad	6416	1.52	1275	0.81	0.71	(0.39)	0.67	(0.71)	0.70	(0.73)	0.61	(0.69)
2	uit de kast	6118	1.50	1048	0.73	0.07	(1.00)	0.45	(0.60)	0.29	(0.75)	0.42	(0.60)
70	in de kast	6118	1.50	1010	0.71	0.55	(0.39)	0.39	(0.67)	0.41	(0.62)	0.34	(0.60)
44	naar de wc	17185	1.94	804	0.61	0.91	(0.36)	1.04	(0.51)	0.93	(0.52)	1.23	(0.48)
62	op de hoek	11205	1.76	756	0.59	0.19	(0.76)	0.14	(0.77)	0.05	(0.99)	0.18	(0.66)

						Time 1				Time 2			
		Freq	Log	Freq	Log	Likert		ME		Likert		ME	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
41	onder elkaar	89055	2.66	688	0.55	0.47	(0.61)	0.45	(0.45)	0.36	(0.64)	0.49	(0.56)
52	in de oren	10856	1.74	667	0.53	-0.36	(0.91)	-0.44	(0.79)	-0.22	(0.89)	-0.54	(0.81)
27	in de oven	2273	1.07	651	0.52	0.60	(0.39)	0.66	(0.59)	0.58	(0.47)	0.57	(0.58)
24	in de pan	4233	1.34	585	0.48	0.43	(0.45)	0.64	(0.62)	0.43	(0.58)	0.46	(0.67)
63	met het doel	26189	2.13	558	0.46	0.24	(0.75)	0.08	(0.94)	0.22	(0.54)	0.23	(0.69)
46	met gemak	3490	1.25	528	0.43	0.58	(0.37)	0.38	(0.63)	0.42	(0.66)	0.56	(0.61)
73	uit het niets	89997	2.66	490	0.40	0.72	(0.42)	0.49	(0.73)	0.53	(0.63)	0.57	(0.54)
57	over de streep	6570	1.53	483	0.39	0.33	(0.56)	0.08	(0.67)	0.33	(0.58)	0.14	(0.66)
58	in het paleis	5394	1.44	427	0.34	-0.15	(0.96)	-0.51	(0.74)	-0.39	(0.93)	-0.40	(0.69)
37	tegen de grond	33283	2.23	369	0.28	-0.28	(0.84)	-0.53	(0.66)	-0.24	(0.72)	-0.69	(0.72)
28	in de bak	5597	1.46	295	0.18	0.13	(0.62)	-0.16	(0.58)	0.02	(0.68)	-0.41	(0.63)
42	van het dak	6202	1.50	280	0.16	-0.40	(0.70)	-0.48	(0.61)	-0.40	(0.78)	-0.43	(0.52)
7	op het boek	74296	2.58	274	0.15	-0.59	(0.83)	-0.64	(0.67)	-0.30	(0.79)	-0.61	(0.55)
39	over de kop	19931	2.01	251	0.11	0.66	(0.36)	0.33	(0.49)	0.40	(0.56)	0.34	(0.54)
75	in de rondte	205	0.02	195	0.00	-0.02	(0.78)	-0.34	(0.70)	-0.28	(0.83)	-0.28	(0.56)

						Time 1				Time 2			
		Freq	Log	Freq	Log	Likert		ME		Likert		ME	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
38	onder de dekens	2585	1.12	175	-0.05	0.59	(0.33)	0.46	(0.69)	0.50	(0.50)	0.49	(0.59)
68	op de film	47205	2.38	145	-0.13	-0.80	(0.87)	-0.78	(0.90)	-0.88	(0.89)	-0.79	(0.84)
33	tegen de vlakte	1682	0.93	141	-0.14	0.19	(0.70)	-0.05	(0.54)	0.13	(0.60)	0.01	(0.67)
17	op de bon	2267	1.06	103	-0.27	0.02	(0.68)	-0.19	(0.74)	0.05	(0.77)	-0.26	(0.72)
6	onder de wol	1068	0.74	96	-0.30	-0.01	(0.86)	-0.07	(0.73)	0.05	(0.73)	0.19	(0.80)
26	uit de kom	803	0.61	95	-0.31	-0.03	(0.67)	-0.16	(0.73)	-0.33	(0.82)	-0.14	(0.58)
53	langs het water	42001	2.33	83	-0.37	0.26	(0.54)	0.07	(0.74)	-0.18	(0.86)	0.04	(0.63)
36	op de pof	93	-0.32	67	-0.46	-0.90	(0.99)	-1.13	(0.84)	-0.78	(0.92)	-0.97	(0.91)
64	over het gras	4481	1.36	54	-0.55	0.14	(0.77)	-0.20	(0.77)	0.01	(0.71)	-0.14	(0.74)
56	tegen mama	8035	1.61	49	-0.59	0.18	(0.72)	0.41	(0.81)	0.37	(0.79)	0.41	(0.88)
40	rond de middag	7659	1.59	48	-0.60	0.60	(0.48)	0.64	(0.53)	0.56	(0.54)	0.66	(0.55)
11	in het geld	59244	2.48	47	-0.61	-1.36	(0.93)	-1.30	(0.47)	-1.30	(0.66)	-1.17	(0.69)
55	in de koffie	12497	1.81	46	-0.62	0.03	(0.89)	0.11	(0.91)	0.16	(0.84)	0.15	(0.90)
31	voor de juf	3250	1.22	29	-0.81	-0.05	(0.71)	-0.47	(0.75)	-0.33	(0.76)	-0.53	(0.68)
8	onder de mat	2443	1.10	28	-0.83	-0.09	(0.75)	-0.28	(0.72)	-0.26	(0.82)	-0.34	(0.89)

						Time 1				Time 2			
		Freq	Log	Freq	Log	Likert		ME		Likert		ME	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
59	uit de kunst	19620	2.00	28	-0.83	-0.23	(0.96)	-0.52	(0.89)	-0.19	(0.91)	-0.47	(0.92)
35	naar de boom	13766	1.85	27	-0.84	-0.58	(0.85)	-0.55	(0.79)	-0.73	(0.70)	-0.75	(0.59)
76	in de foto	35457	2.26	26	-0.86	-1.40	(0.84)	-1.19	(0.68)	-1.30	(1.06)	-1.23	(0.76)
48	naast de auto	37927	2.29	25	-0.88	0.15	(0.70)	-0.05	(0.66)	-0.10	(0.73)	-0.04	(0.79)
34	uit de gang	17176	1.94	24	-0.89	-0.94	(0.84)	-0.93	(0.61)	-0.79	(0.91)	-0.99	(0.60)
72	langs de tafel	17185	1.94	22	-0.93	-0.36	(0.84)	-0.66	(0.72)	-0.72	(0.95)	-0.44	(0.64)
9	onder het asfalt	1208	0.79	13	-1.15	-1.40	(0.98)	-0.98	(0.67)	-1.21	(0.87)	-1.19	(0.63)
67	met de schoen	7970	1.61	11	-1.21	-0.89	(0.95)	-0.88	(0.52)	-0.83	(0.80)	-0.84	(0.67)
50	om de tekeningen	5756	1.47	9	-1.29	-1.48	(0.74)	-1.24	(0.60)	-1.22	(0.97)	-1.20	(0.48)
69	op de meester	7159	1.56	8	-1.34	-1.51	(0.88)	-1.45	(0.54)	-1.71	(0.86)	-1.53	(0.53)
78	rond de ingang	8174	1.62	8	-1.34	-0.81	(0.79)	-0.88	(0.75)	-0.71	(0.84)	-0.69	(0.68)
79	uit de trommel	716	0.56	8	-1.34	-0.47	(1.00)	-0.83	(0.69)	-0.61	(0.84)	-0.80	(0.57)

						Time 1				Time 2			
		Freq	Log	Freq	Log	Likert		ME		Likert		ME	
		N	FreqN	PP	FreqPP	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
3	bij de fietsen	8807	1.65	6	-1.45	-0.31	(1.02)	0.20	(0.79)	0.30	(0.62)	-0.05	(0.94)
5	in de groente	4882	1.40	6	-1.45	-1.08	(0.84)	-1.03	(0.53)	-0.83	(0.83)	-1.08	(0.65)
29	in de piano	3534	1.26	6	-1.45	-1.54	(0.67)	-1.33	(0.52)	-1.52	(0.61)	-1.40	(0.42)
12	langs de auto	37927	2.29	5	-1.51	-0.21	(0.81)	-0.26	(0.78)	-0.28	(0.87)	-0.17	(0.68)
32	naast het café	7456	1.58	5	-1.51	0.16	(0.69)	0.05	(0.64)	0.24	(0.65)	-0.07	(0.69)
45	lans het park	9253	1.67	4	-1.59	-0.59	(0.83)	-0.60	(0.70)	-0.58	(0.84)	-0.36	(0.64)
10	in de shampoo	458	0.37	3	-1.69	-0.95	(0.95)	-1.02	(0.66)	-0.84	(0.74)	-1.05	(0.74)
18	naast het hek	3778	1.29	3	-1.69	-0.16	(0.64)	-0.10	(0.67)	-0.06	(0.67)	-0.20	(0.63)
20	langs de boeken	8777	1.65	3	-1.69	-1.13	(1.00)	-1.08	(0.58)	-1.12	(0.68)	-0.97	(0.48)
30	naast de bloemen	9294	1.68	2	-1.81	-0.48	(0.83)	-0.57	(0.77)	-0.43	(0.69)	-0.78	(0.60)
19	voor de schommel	274	0.15	1	-1.99	-0.79	(0.82)	-0.65	(0.72)	-0.48	(0.84)	-0.84	(0.56)
65	over het karton	603	0.49	1	-1.99	-1.43	(0.69)	-1.35	(0.44)	-1.43	(0.75)	-1.32	(0.58)

Appendix 3.3 Linear mixed-effects models

We fitted linear mixed-effects models (Baayen et al. 2008), using the LMER function from the lme4 package in R (version 3.2.3; CRAN project; R Core Team, 2015), first to the familiarity judgments and then to the Δ-scores.

In the first analysis, we investigated to what extent the familiarity judgments can be predicted by the frequency of the specific phrase (LogFreqPP) and the lemma-frequency of the noun (LogFreqN), and to what degree the factors RatingScale (0 = Likert, 1 = Magnitude Estimation) and Time (0 = first session, 1 = second session) exert influence. The fixed effects were standardized. Participants and items were included as random effects. We incorporated a random intercept for items and random slopes for both items and participants to account for between-item and between-participant variation. The model does not contain a by-participant random intercept, because after the Z-score transformation all participants' scores have a mean of 0 and a standard deviation of 1.

We started with a random intercept only model. We added fixed effects, and all two-way interactions, one by one and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in familiarity judgments. We started with LogFreqPP ($\chi^2(1) = 86.64$, $p < .001$). After that, we added LogFreqN ($\chi^2(1) = 0.03$, $p = .87$) and the interaction term LogFreqPP x LogFreqN ($\chi^2(1) = 0.002$, $p = .96$), which did not improve model fit. We then proceeded with RatingScale ($\chi^2(1) = 0.0003$, $p = .99$), which did not improve model fit either. The interaction term RatingScale x LogFreqPP did contribute to the fit of the model ($\chi^2(2) = 21.79$, $p < .001$), as did RatingScale x LogFreqN ($\chi^2(2) = 6.77$, $p < .05$). There cannot be a main effect of Time in this analysis, since scores were converted to Z-scores for the two sessions separately (i.e. the mean scores at Time 1 and Time 2 were 0). We did include the two-way interactions of Time and the other factors. None of these was found to improve model fit (Time x RatingScale ($\chi^2(2) = 0.00$, $p = .99$); Time x LogFreqPP ($\chi^2(1) = 0.01$, $p = .91$); Time x LogFreqN ($\chi^2(1) = 0.01$, $p = .91$)). Finally, PresentationOrder did not contribute to the goodness-of-fit ($\chi^2(1) = 1.27$, $p = .26$). Apart from the interaction term PresentationOrder x RatingScale ($\chi^2(2) = 7.05$, $p = .03$), none of the interactions of PresentationOrder and the other predictors in the model was found to improve model fit (PresentationOrder x LogFreqPP ($\chi^2(1) = 1.89$, $p = .17$); PresentationOrder x LogFreqN ($\chi^2(1) = 0.38$, $p = .54$); PresentationOrder x Time ($\chi^2(1) = 1.27$, $p = .26$); PresentationOrder x LogFreqPP x RatingScale ($\chi^2(2) = 5.41$, $p = .07$); PresentationOrder x LogFreqN x RatingScale ($\chi^2(2) = 0.46$, $p = .80$)). The model selection procedure thus resulted in a model comprising

LogFreqPP, LogFreqN, RatingScale, RatingScale x LogFreqPP, RatingScale x LogFreqN, and PresentationOrder x Ratingscale.

We then added a by-item random slope for RatingScale and by-participant random slopes for LogFreqPP and LogFreqN. There are no by-item random slopes for the factors LogFreqPP, LogFreqN, PresentationOrder, and the interactions involving these factors, because each item has only one phrase frequency, one lemma frequency, and a fixed position in the order of presentation. There is no by-participant random slope for RatingScale, since half of the participants only used one scale. Within these limits, a model with a full random effect structure was constructed following Barr et al. (2013). Subsequently, we excluded random slopes with the lowest variance step by step until a further reduction would imply a significant loss in the goodness of fit of the model (Matuschek et al. 2017). Model comparisons indicated that the inclusion of the by-participant random slopes for LogFreqPP, LogFreqN, and PresentationOrder, and the by-item random slope for RatingScale was justified by the data ($\chi2(3) = 90.21$, $p < .001$). Inspection of the variance inflation factors revealed that there do not appear to be harmful effects of collinearity (the highest VIF value is 1.20; tolerance statistics are 0.83 or more, cf. Field et al. 2012: 275). Confidence intervals were estimated via parametric bootstrapping over 1000 iterations (Bates et al. 2015). The model is summarized in Table 3.2.

In a separate analysis, we ran linear mixed-effects models on the Δ-scores, to determine which factors influence variation across time. The absolute Δ-scores indicate the extent to which a participant's rating for a particular item at Time 2 differs from the rating at Time 1 (see Section 3.3.5). For each item, we have a list of 91 Δ-scores that express each participant's stability in the grading. In order to fit a linear mixed-effects model on the set of Δ-scores, we log-transformed them using the natural logarithm function. The absolute Δ-scores constitute the positive half of a normal distribution. Log-transforming the scores yields a normal distribution, thus complying with the assumptions of parametric statistical tests.

LogFreqPP, LogFreqN, RatingScaleT1 and RatingScaleT2 (the type of scale used at Time 1 and Time 2 respectively, i.e. Likert or ME), and PresentationOrder were included as fixed effects and standardized. Participants and items were included as random effects. We incorporated a random intercept for both items and participants to account for between-item and between-participant variation. We then added fixed effects one by one and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in log-transformed absolute Δ-scores. We started with LogFreqPP ($\chi^2(1) = 32.92$, $p < .001$). After that, we added LogFreqN ($\chi^2(1) = 0.04$, $p = .84$). Given that LogFreqN did not improve model fit, we left out this predictor. We then proceeded with RatingScaleT1 ($\chi^2(1) = 0.15$, $p = .70$) and RatingScaleT2 ($\chi^2(1) = 2.39$, $p = .12$), neither of which improved model fit. The interaction term RatingScaleT1 x RatingScaleT2 did not contribute to the fit of the model fit either (χ^2(3) = 6.67, *p* = .08). The interaction term RatingScaleT1 x LogFreqPP did improve model fit (χ^2(2) = 40.94, *p* < .001), as did RatingScaleT2 x LogFreqPP (χ^2(2) = 13.91, *p* < .001). The three-way interaction RatingScaleT1 x RatingScaleT2 x LogFreqPP did not explain a significant portion of variance (χ^2(2) = 4.63, *p* = .10). Finally, neither PresentationOrder (χ^2(1) = 0.27, *p* = .60), nor any of the interactions of PresentationOrder and the other predictors in the model was found to improve model fit (PresentationOrder x LogFreqPP (χ^2(1) = 1.75, *p* = .19); PresentationOrder x LogFreqPP x RatingScaleT1 (χ^2(2) = 2.52, *p* = .28); PresentationOrder x LogFreqPP x RatingScaleT2 (χ^2(2) = 1.78, *p* = .41)). The model selection procedure thus resulted in a model comprising LogFreqPP, RatingScaleT1 x LogFreqPP, and RatingScaleT2 x LogFreqPP.

We then added by-item random slopes for RatingScaleT1 and RatingScaleT2, and a by-participant random slope for LogFreqPP, thus constructing a model with a full random effect structure following Barr et al. (2013). Subsequently, we excluded random slopes with the lowest variance step by step until a further reduction would imply a significant loss in the goodness of fit of the model (Matuschek et al. 2017). Model comparisons indicated that the inclusion of the by-item random slope for RatingScaleT1 and the by-participant random slopes for LogFreqPP was justified by the data (χ2(2) = 12.96, *p* < .01). Inspection of the variance inflation factors revealed that there do not appear to be harmful effects of collinearity (the highest VIF value is 2.76; tolerance statistics are 0.36 or more). Again, confidence intervals were estimated via parametric bootstrapping over 1000 iterations. The model is summarized in Table 3.5.

Appendix 4.1 Job ad word sequences and corpus-based frequencies and surprisal estimates

The Job ad word sequences; base-10 logarithm of the frequency of occurrence per million words in the Job ad corpus and the NLCOW14-subset for the phrase as a whole and for the final word (lemma search); the surprisal of the final word based on data in NLCOW14-subset.

		Based on Job ad corpus	Based on NLCOW14-subset		
		LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
1	40 uur per week	2.52	-0.40	41	0.92
2	voor meer informatie	2.36	0.37	84	1.33
3	kennis en ervaring	2.10	0.12	110	1.07
4	hoog in het vaandel	1.84	0.34	24	-0.32
5	werving en selectie	1.82	-0.54	119	0.63
6	een vast dienstverband	1.87	-0.77	332	-0.09
7	voor langere tijd	1.65	0.08	91	1.33
8	het eerste aanspreekpunt	1.48	-0.63	397	-0.15
9	goede contactuele eigenschappen	1.39	-1.22	339	0.82
10	bij gebleken geschiktheid	1.32	-1.06	217	-0.24

		Based on Job ad corpus	Based on NLCOW14-subset		
		LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
11	academisch werk- en denkniveau	1.00	-1.33	29	-0.85
12	een grote mate van zelfstandigheid	1.15	-0.99	46	0.07
13	in een hecht team	0.82	-1.57	119	1.08
14	een persoonlijk ontwikkelingsplan	0.55	-1.27	537	-0.71
15	een sterk analytisch vermogen	0.67	-1.69	208	0.89
16	met de mogelijkheid tot verlenging	0.50	-1.69	68	0.17
17	in de breedste zin van het woord	0.94	-0.04	9	0.96
18	met een afstand tot de arbeidsmarkt	0.05	-1.06	20	0.17
19	het geschetste profiel	0.24	-1.87	1546	0.58
20	in de meest uiteenlopende sectoren	0.39	-2.17	135	0.34
21	een vliegende start	0.10	-0.49	226	1.23
22	bewijs van goed gedrag	0.11	-0.99	71	1.17
23	conform de geldende CAO	-0.08	-1.87	151	0.20
24	met behoud van uitkering	-0.02	-0.51	45	0.47
25	bevoegd en bekwaam	-0.08	-1.39	247	0.04

		Based on Job ad corpus	Based on NLCOW14-subset		
		LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
26	een integrale benadering	-0.17	-0.81	342	0.83
27	naar aanleiding van de advertentie	-0.56	-2.17	96	0.28
28	eenvoudige administratieve werkzaamheden	-0.51	-1.87	919	0.82
29	een scherpe blik	-0.52	-1.17	447	0.86
30	buiten de geijkte paden	-0.90	-1.57	110	0.37
31	affiniteit met het onderwerp	-0.74	-1.69	112	0.84
32	een internationale speler van formaat	-1.24	-2.17	519	0.55
33	een flinke portie lef	-1.39	-2.17	344	0.07
34	met bewezen kwaliteiten	-1.17	-2.17	1586	0.59
35	een collegiale opstelling	-1.29	-2.17	13960	0.55

Appendix 4.2 News report word sequences and corpus-based frequencies and surprisal estimates

The News report word sequences; base-10 logarithm of the frequency of occurrence per million words in the Twente News Corpus and the NLCOW14-subset for the phrase as a whole and for the final word (lemma search); the surprisal of the final word based on data in NLCOW14-subset.

		Based on News report corpus	Based on NLCOW14-subset		
		LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
36	de Tweede Kamer	1.94	0.38	144	0.31
37	wetenschap en techniek	1.87	-0.67	211	0.81
38	verkeer en vervoer	1.80	-0.52	169	0.57
39	in elk geval	1.71	0.84	52	0.65
40	in de Verenigde Staten	1.66	0.82	27	0.20
41	het openbaar ministerie	1.16	-0.32	264	0.22
42	de negentiende eeuw	1.05	-0.58	269	0.42
43	de raad van bestuur	1.04	-2.17	101	0.77
44	aan de andere kant	1.22	0.81	28	1.10
45	evenementen en manifestaties	1.47	-2.17	4662	0.00

		Based on News report corpus	Based on NLCOW14-subset		
		LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
46	het dagelijks leven	0.97	-0.20	213	1.50
47	op een gegeven moment	0.98	0.54	32	0.85
48	met terugwerkende kracht	0.58	0.26	77	1.03
49	in volle gang	0.66	0.05	96	0.60
50	een doorn in het oog	0.55	0.01	19	0.76
51	op geen enkele wijze	0.19	0.15	30	1.23
52	aan het begin van het seizoen	0.00	-0.38	15	0.74
53	de lokale bevolking	0.30	-0.25	329	0.78
54	het centrum van de stad	0.42	-0.59	50	1.02
55	correcties en aanvullingen	0.32	-1.17	189	0.06
56	de opvang van asielzoekers	-0.05	-1.09	149	0.32
57	de traditionele partijen	-0.42	-1.06	617	0.97
58	op last van de rechter	-0.35	-2.17	27	0.60
59	in de huidige situatie	-0.20	-0.22	56	0.99
60	een onafhankelijke commissie	-0.09	-0.83	382	0.65

		Based on News report corpus	Based on NLCOW14-subset		
		LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
61	een criminele afrekening	-0.83	-1.87	1486	-0.13
62	de koninklijke loge	-0.90	-1.87	1318	-0.54
63	een ingrijpende herstructurering	-0.90	-1.69	895	-0.01
64	op weg naar de top	-0.71	-1.22	32	0.93
65	in het belang van het kind	-0.63	-0.57	16	1.01
66	aan de vooravond van een revolutie	-1.36	-1.87	46	0.40
67	de uitkomsten van het rapport	-1.30	-1.69	73	0.78
68	met hernieuwde energie	-1.46	-1.17	262	1.03
69	een ongekende vrijheid	-1.38	-2.17	886	0.79
70	een luxe jacht	-1.46	-2.17	1092	0.35

Appendix 4.3 Average Stereotypy Scores for the Job ad stimuli

		Stereotypy Scores					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
1	40 uur per	97.5	(15.8)	97.5	(15.8)	90.5	(29.7)
2	voor meer	58.2	(48.1)	58.3	(48.1)	55.4	(48.6)
3	kennis en	21.0	(29.2)	12.9	(23.7)	6.3	(19.1)
4	hoog in het	90.0	(30.4)	82.5	(38.5)	66.7	(47.7)
5	werving en	96.7	(0.0)	84.6	(32.4)	27.6	(44.2)
6	een vast	15.8	(19.8)	13.7	(22.0)	5.2	(8.3)
7	voor langere	47.9	(43.7)	64.9	(38.0)	63.3	(40.4)
8	het eerste	2.4	(13.0)	0.2	(0.6)	0.0	(0.1)
9	goede contactuele	57.5	(50.1)	52.5	(50.6)	2.4	(15.4)
10	bij gebleken	74.8	(43.7)	29.9	(46.3)	2.4	(15.4)
11	academisch werk- en	85.0	(36.2)	57.5	(50.1)	0.0	(0.0)
12	een grote mate van	25.4	(32.2)	11.8	(25.5)	3.2	(14.3)
13	in een hecht	52.5	(50.6)	40.0	(49.6)	11.9	(32.8)

		Stereotypy Scores					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
14	een persoonlijk	17.3	(23.4)	13.7	(21.9)	13.1	(21.5)
15	een sterk analytisch	95.0	(22.1)	80.0	(40.5)	66.7	(47.7)
16	met de mogelijkheid tot	33.9	(46.0)	22.0	(40.3)	1.0	(1.9)
17	in de breedste zin van het	100.0	(0.0)	95.0	(22.1)	78.6	(41.5)
18	met een afstand tot de	55.0	(50.4)	2.5	(15.8)	0.0	(0.0)
19	het geschetste	35.0	(48.3)	17.5	(38.5)	0.0	(0.0)
20	in de meest uiteenlopende	0.2	(0.4)	0.1	(0.3)	0.2	(0.4)
21	een vliegende	77.8	(38.3)	70.5	(43.2)	13.9	(34.2)
22	bewijs van goed	97.5	(15.8)	100.0	(0.0)	76.2	(43.1)
23	conform de geldende	13.9	(31.8)	5.3	(15.6)	4.1	(2.5)
24	met behoud van	25.8	(32.0)	9.7	(23.3)	0.0	(0.0)
25	bevoegd en	22.5	(42.3)	17.5	(38.5)	7.1	(26.1)

		Stereotypy Scores					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
26	een integrale	12.9	(20.5)	13.4	(21.1)	0.2	(1.1)
27	naar aanleiding van de	9.4	(22.6)	6.5	(19.0)	0.0	(0.0)
28	eenvoudige administratieve	50.0	(50.6)	50.0	(50.6)	28.6	(45.7)
29	een scherpe	9.0	(9.7)	9.4	(8.5)	5.1	(8.2)
30	buiten de geijkte	56.1	(36.4)	48.6	(40.8)	4.3	(17.8)
31	affiniteit met het	13.4	(17.7)	10.3	(17.5)	8.8	(15.4)
32	een internationale speler van	2.5	(15.8)	7.5	(26.7)	0.0	(0.0)
33	een flinke portie	0.0	(0.0)	0.9	(5.6)	0.8	(5.4)
34	met bewezen	2.5	(15.8)	2.5	(15.8)	2.4	(15.4)
35	een collegiale	14.1	(33.6)	11.8	(31.2)	0.0	(0.0)

Appendix 4.4 Average Stereotypy Scores for the News report stimuli

		Stereotypy Scores					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
36	de Tweede	34.9	(18.2)	39.4	(16.9)	32.6	(17.2)
37	wetenschap en	5.3	(20.5)	8.0	(23.0)	10.6	(28.3)
38	verkeer en	7.5	(21.0)	15.9	(25.5)	2.7	(7.5)
39	in elk	73.1	(42.7)	87.7	(29.6)	65.0	(46.4)
40	in de Verenigde	86.5	(32.9)	96.3	(15.5)	94.1	(21.3)
41	het openbaar	21.9	(36.2)	22.3	(36.6)	14.7	(31.2)
42	de negentiende	72.8	(42.6)	63.1	(46.9)	50.8	(49.0)
43	de raad van	30.7	(13.8)	27.5	(18.5)	24.0	(18.9)
44	aan de andere	98.0	(0.6)	95.7	(15.5)	98.3	(0.9)
45	evenementen en	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)
46	het dagelijks	34.4	(29.9)	46.1	(25.9)	38.5	(31.8)
47	op een gegeven	100.0	(0.0)	95.0	(22.1)	100.0	(0.0)
48	met terugwerkende	97.5	(15.8)	97.5	(15.8)	92.9	(26.1)
49	in volle	11.4	(21.1)	10.0	(18.5)	7.9	(14.2)

		Stereotypy Scores					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
50	een doorn in het	95.0	(22.1)	97.5	(15.8)	90.5	(29.7)
51	op geen enkele	53.7	(31.4)	61.9	(29.2)	60.7	(32.7)
52	aan het begin van het	3.6	(12.5)	5.7	(13.5)	7.5	(19.2)
53	de lokale	7.0	(12.5)	5.5	(10.3)	7.8	(12.0)
54	het centrum van de	54.7	(47.6)	45.2	(48.1)	68.0	(43.5)
55	correcties en	22.5	(42.3)	25.0	(43.9)	7.1	(26.1)
56	de opvang van	8.3	(24.3)	4.3	(17.7)	6.6	(20.8)
57	de traditionele	3.6	(7.6)	4.1	(6.9)	1.2	(3.8)
58	op last van de	0.0	(0.0)	7.5	(26.7)	9.5	(29.7)
59	in de huidige	19.2	(17.1)	12.8	(16.1)	16.5	(16.4)

		Stereotypy Scores					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
60	een onafhankelijke	5.2	(10.9)	3.4	(8.6)	7.0	(12.1)
61	een criminele	28.8	(44.5)	26.4	(43.3)	13.7	(33.9)
62	de koninklijke	13.6	(13.2)	11.6	(12.9)	17.8	(13.9)
63	een ingrijpende	5.5	(4.2)	6.3	(5.7)	4.8	(4.2)
64	op weg naar de	3.9	(13.9)	11.7	(24.7)	1.3	(8.5)
65	in het belang van het	3.0	(12.0)	1.9	(10.9)	1.6	(10.6)
66	aan de vooravond van een	0.0	(0.0)	0.0	(0.0)	0.0	(0.0)
67	de uitkomsten van het	63.6	(44.3)	77.4	(36.1)	62.5	(44.7)
68	met hernieuwde	12.5	(33.5)	10.0	(30.4)	2.4	(15.4)
69	een ongekende	2.2	(9.7)	1.4	(8.9)	0.0	(0.0)
70	een luxe	8.3	(8.8)	14.4	(13.8)	12.5	(18.1)

Appendix 4.5 Mixed-effects logistic regression model fitted to the completion task data

The stereotypy scores were not normally distributed. Therefore, it was not justified to fit a linear mixed-effects model. We used a mixed-effects logistic regression model (Jaeger 2008) instead. Per response, we indicated whether or not it corresponded to a complement observed in the specialized corpora. By means of a mixed logit-model, we investigated whether there are significant differences across groups of participants and/or sets of stimuli in the proportion of responses that correspond to a complement in the specialized corpora. We fitted this model using the LMER function from the lme4 package in R (version 3.3.3; CRAN project; R Core Team, 2017). Group, Itemtype, and their interaction were included as fixed effects, and participants and items as random effects. The fixed effects were standardized. Random intercepts and random slopes for participants and items were included to account for between-subject and between-item variation.[33]

A model with a full random effect structure was constructed following Barr, Levy, Scheepers, and Tily (2013). A comparison with the intercept-only model proved that the inclusion of the by-item random slope for Group and the by-participant random slope for Itemtype was justified by the data ($\chi^2(7) = 174.83$, $p < .001$). Confidence intervals were estimated via parametric bootstrapping over 1000 iterations (Bates, Mächler, Bolker & Walker 2015).

In order to obtain all relevant comparisons of the three groups and the two types of stimuli, we ran the model with different coding schemes and we report 99% confidence intervals (as opposed to the more common 95%) to correct for multiple comparisons. Since the groups were not expected to differ systematically in experience with News report word sequences, none of the groups forms a natural baseline in this respect. As for the Job ad stimuli, from a usage-based perspective, differences between Recruiters and Job-seekers are as interesting as differences between Job-seekers and Inexperienced participants, or Recruiters and Inexperienced participants. Therefore, we treatment-coded the factors, first using *Recruiters* as the reference group for Group and *Job ad stimuli* as the reference group for Itemtype. The resulting model is summarized in Table 4.6. The intercept represents the proportion of the Recruiters' responses to the Job ad stimuli that correspond to a complement in the Job ad corpus. This proportion does not differ significantly from the proportion of their responses to the News report items that correspond to a complement in the Twente News Corpus.

[33] By-participant random slopes for Group were not included, as this was a between-participants factor; by-item random slopes for Itemtype were not included, as this was a between-items factor.

There are significant differences between the groups of participants on the Job ad stimuli. Both the Inexperienced participants and the Job-seekers have significantly lower proportions of responses to the Job ad stimuli that match a complement in the Job ad corpus than the Recruiters. The model also reveals that the difference between the proportions on the two types of stimuli is significantly different across groups.

Table 4.6 Mixed-effects logistic regression model (family: binomial) fitted to the responses to the completion task (0 = does not correspond to a complement in the specialized corpus; 1 = corresponds to a complement in the specialized corpus), using *Recruiters–Job ad stimuli* as the reference condition.

	Estimate	*SE*	*z*	99 % CI	
(Intercept)	0.56	0.43	1.31	*-0.54, 1.65*	
Itemtype_NewsReport	-0.56	0.60	-0.93	*-2.06, 0.97*	
Group_Jobseekers	-0.69	0.17	-4.09	*-1.11, -0.26*	**
Group_Inexperienced	-2.38	0.29	-8.30	*-3.11, -1.64*	**
Itemtype_NewsReport x Group_ Jobseekers	0.91	0.21	4.36	*0.36, 1.46*	**
Itemtype_ NewsReport x Group_Inexperienced	2.14	0.38	5.62	*1.15, 3.09*	**

Note: Significance code: 0.01 ‘**’

To examine the remaining differences, we then used *Job-seekers–Job ad stimuli* as the reference condition. The outcomes are summarized in Table 4.7. The proportion of the Job-seekers' responses to the Job ad items that correspond to a complement in the Job ad corpus does not differ significantly from the proportion of their responses to the News report items that match a complement in the Twente News Corpus. Furthermore, the outcomes show that the Job-seekers' responses to the Job ad stimuli were significantly more likely to correspond to a complement in the Job ad corpus than the responses of the Inexperienced participants. In addition, the model reveals that the difference between the proportions on the two types of stimuli is significantly different for the Inexperienced participants compared to the Job-seekers.

Table 4.7 Mixed-effects logistic regression model (family: binomial) fitted to the responses to the completion task (0 = does not correspond to a complement in the specialized corpus; 1 = corresponds to a complement in the specialized corpus), using *Job-seekers–Job ad stimuli* as the reference condition.

	Estimate	*SE*	*z*	99 % CI	
(Intercept)	-0.13	0.40	-0.32	*-1.13, 0.86*	
Itemtype_NewsReport	0.35	0.56	0.63	*-1.04, 1.72*	
Group_ Inexperienced	-1.69	0.25	-6.78	*-2.34, -1.04*	**
Group_Recruiters	0.69	0.17	4.09	*0.25, 1.14*	**
Itemtype_NewsReport x Group_ Inexper.	1.23	0.32	3.79	*0.38, 2.07*	**
Itemtype_ NewsReport x Group_ Recruiters	-0.91	0.21	-4.36	*-1.43, - 0.39*	**

Note: Significance code: 0.01 '**'

Finally, we used *Inexperienced-News report stimuli* as the reference condition. The outcomes, summarized in Table 4.8, show that the proportion of the Inexperienced participants' responses to the Job ad items that correspond to a complement in the specialized corpus is not significantly different from the proportion of their responses to the News report items that match a complement in the specialized corpus. They also reveal that the three groups do not differ significantly from each other in the proportion of responses to the News report stimuli that match a complement in the specialized corpus.

Table 4.8 Mixed-effects logistic regression model (family: binomial) fitted to the responses to the completion task (0 = does not correspond to a complement in the specialized corpus; 1 = corresponds to a complement in the specialized corpus), using *Inexperienced−News report stimuli* as the reference condition.

	Estimate	*SE*	*z*	99 % CI	
(Intercept)	-0.24	0.45	-0.62	*-1.34, 0.84*	
Itemtype_JobAd	-1.58	0.64	-2.47	*-3.12, 0.01*	
Group_ Jobseekers	0.46	0.23	1.98	*-0.11, 1.04*	
Group_Recruiters	0.24	0.27	0.88	*-0.44, 0.92*	
Itemtype_ JobAd x Group_ Jobseekers	1.23	0.32	3.79	*0.38, 2.04*	**
Itemtype_ JobAd x Group_ Recruiters	2.14	0.38	5.61	*1.14, 3.11*	**

Note: Significance code: 0.01 '**'

Appendix 4.6 Linear mixed-effects models fitted to the voice onset times (VOT task)

We fitted linear mixed-effects models (Baayen et al. 2008), using the LMER function from the lme4 package in R (version 3.3.2; CRAN project; R Core Team 2017), to the Voice Onset Times. First, we investigated whether there are significant differences in VOTs across groups of participants and/or sets of stimuli, similar to our analysis of the stereotypy scores. Subsequently, we examined to what extent the VOTs can be predicted by word length, corpus-based word frequency, presentation order, and different measures of word predictability.

In the first analysis, Group, Itemtype, and their interaction were included as fixed effects, and participants and items as random effects. The fixed effects were standardized. We included random intercepts and slopes for participants and items to account for between-subject and between-item variation.[34]

A model with a full random effect structure was constructed following Barr et al. (2013). A comparison with the intercept-only model proved that the inclusion of the by-item random slope for Group and the by-participant random slope for Itemtype was justified by the data ($\chi^2(7) = 34.34$, $p < .001$). The variance explained by this model is 59% ($R^2m = .04$, $R^2c = .59$).[35] Confidence intervals were estimated via parametric bootstrapping over 1000 iterations (Bates et al. 2015).

In order to obtain all relevant comparisons of the three groups and the two types of stimuli, we ran the model with different coding schemes and we report 99% confidence intervals to correct for multiple comparisons. We treatment-coded the factors, first using *Recruiters* as the reference group for Group and *Job ad stimuli* as the reference group for Itemtype. The resulting model is summarized in Table 4.9. The intercept represents the mean VOT of the Recruiters on the Job ad stimuli. Subsequently, we used *Job-seekers–Job ad stimuli* as the reference condition (Table 4.10), and finally *Inexperienced-News report stimuli* (Table 4.11).

The models reveal that none of the groups shows a significant difference between VOTs on the News report items and VOTs on the Job ad items. The Inexperienced do differ significantly from the Recruiters and Job-seekers in the relationship between the two sets of items. The majority of the Recruiters and the Job-seekers responded faster to the Job ad items than to the News report items (as evidenced by the Recruiters' and Job-seekers' marks below the zero line in

[34] By-participant random slopes for Group were not included, as this was a between-participants factor; by-item random slopes for Itemtype were not included, as this was a between-items factor.

[35] R^2m (marginal R^2 coefficient) represents the amount of variance explained by the fixed effects; R^2c (conditional R^2 coefficient) is interpreted as variance explained by both fixed and random effects (i.e. the full model) (Johnson 2014).

Figure 4.4). For the vast majority of the Inexperienced participants it is just the other way around: they were faster on the News report stimuli compared to the Job add stimuli. The mixed-effects models indicate that the Inexperienced participants' data pattern is significantly different from the Recruiters' and the Job-seekers'.

Table 4.9 Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using *Recruiters–Job ad stimuli* as the reference condition.

	Estimate	*SE*	*t*	99 % CI	
(Intercept)	0.522	0.017	30.27	*0.477, 0.566*	
Itemtype_NewsReport	0.020	0.016	1.24	*-0.024, 0.064*	
Group_Jobseekers	0.009	0.019	0.50	*-0.036, 0.057*	
Group_Inexperienced	-0.036	0.019	-1.93	*-0.085, 0.013*	
Itemtype_NewsReport x Group_ Jobseekers	-0.011	0.006	-1.88	*-0.026, 0.004*	
Itemtype_ NewsReport x Group_Inexperienced	-0.030	0.007	-4.12	*-0.048, -0.011*	**

Note: Significance code: 0.01 '**'

Table 4.10 Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using *Job-seekers–Job ad stimuli* as the reference condition.

	Estimate	*SE*	*t*	99 % CI	
(Intercept)	0.531	0.017	32.09	*0.488, 0.574*	
Itemtype_NewsReport	0.009	0.015	0.62	*-0.028, 0.047*	
Group_ Recruiters	-0.009	0.019	-0.50	*-0.058, 0.040*	
Group_Inexperienced	-0.045	0.018	-2.47	*-0.094, 0.003*	
Itemtype_NewsReport x Group_ Recruiters	0.011	0.006	1.88	*-0.004, 0.026*	
Itemtype_NewsReport x Group_ Inexperienced	-0.019	0.005	-3.43	*-0.034, -0.004*	**

Note: Significance code: 0.01 '**'

Table 4.11 Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using *Inexperienced–News report stimuli* as the reference condition.

	Estimate	*SE*	*t*	99 % CI	
(Intercept)	0.476	0.017	28.70	*0.434, 0.520*	
Itemtype_JobAd	0.010	0.016	0.61	*-0.031, 0.048*	
Group_ Recruiters	0.066	0.018	3.56	*0.016, 0.115*	**
Group_ Jobseekers	0.064	0.018	3.53	*0.017, 0.111*	**
Itemtype_ JobAd x Group_ Recruiters	-0.030	0.007	-4.12	*-0.048, -0.011*	**
Itemtype_ JobAd x Group_ Jobseekers	-0.019	0.005	-3.43	*-0.033, -0.005*	**

Note: Significance code: 0.01 '**'

In the second analysis, we investigated to what extent the VOTs can be predicted by various characteristics of the target words. We included the length of the target word in letters (WordLength), and its lemma-frequency, residualized against word length (rLogFreq), as they are known to affect naming times. In addition, we examined possible effects of PresentationOrder and block, as artifacts of our experimental design. Furthermore, we investigated three different operationalizations of word predictability. GenericSurprisal is the surprisal of the target word given the cue, estimated by language models trained on the generic corpus meant to reflect Dutch readers' overall experience. ClozeProbability amounts to the percentage of participants that complemented the cue with the target word in the completion task preceding the VOT task. The binary variable TargetMentioned indicates whether or not the target word had been mentioned by a given participant in the completion task. The fixed effects were standardized. Participants and items were included as random effects. We incorporated a random intercept for both items and participants to account for between-item and between-participant variation. We then added fixed effects one by one and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in voice onset times.

We started with WordLength ($\chi^2(1) = 13.73$, $p < .001$), followed by rLogFreq ($\chi^2(1) = 4.78$, $p < .05$), and PresentationOrder ($\chi^2(1) = 3.97$, $p < .05$). After that, we added block ($\chi^2(1) = 2.10$, $p = .15$) and the interaction term PresentationOrder x block ($\chi^2(1) = 0.01$, $p = .93$). Given that neither of the latter two improved model fit, we left out these predictors. We then proceeded with the predictability measures, starting with the most general one: GenericSurprisal. This predictor did not contribute to the fit of the model ($\chi^2(1) = 2.54$, $p = .11$) and therefore we omitted it. ClozeProbability did improve model fit ($\chi^2(1) = 49.22$, $p < .001$), as did TargetMentioned ($\chi^2(1) = 309.37$, $p < .001$). We then included the interaction term rLogFreq x ClozeProbability, which did not contribute to the fit of the model fit ($\chi^2(1) = 3.60$, $p = .06$). rLogFreq x TargetMentioned did explain a significant portion of variance ($\chi^2(1) = 16.75$, $p < .001$). Finally, none of the two-way interactions of PresentationOrder and the other predictors in the model was found to improve model fit (PresentationOrder x TargetMentioned ($\chi^2(1) = 0.57$, $p = .45$); PresentationOrder x ClozeProbability ($\chi^2(1) = 0.65$, $p = .42$); PresentationOrder x rLogFreq ($\chi^2(1) = 0.21$, $p = .65$); PresentationOrder x WordLength ($\chi^2(1) = 2.58$, $p = .11$)). The model selection procedure thus resulted in a model comprising WordLength, rLogFreq, PresentationOrder, ClozeProbability, TargetMentioned, and rLogFreq x TargetMentioned.

We then added random slopes for participants. There are no by-item random slopes, because each item has only one lemma frequency, one cloze probability, one corpus-based surprisal estimate, one length, and a fixed position in the presentation order. Furthermore, there are items no one had mentioned in the completion task, thus prohibiting by-item random slopes for TargetMentioned. Within these limits, a model with a full random effect structure was constructed following Barr et al. (2013). Subsequently, we excluded random slopes with the lowest variance step by step until a further reduction would imply a significant loss in the goodness of fit of the model (Matuschek et al. 2017). Model comparisons indicated that the inclusion of the by-participant random slopes for WordLength, PresentationOrder, ClozeProbability, and TargetMentioned was justified by the data ($\chi 2(5) = 53.00$, $p < .001$). Then, confidence intervals were estimated via parametric bootstrapping over 1000 iterations (Bates et al. 2015). We first ran the model using *Target not mentioned* as the reference condition and then *Target mentioned*. The outcomes are presented in Table 4.5 in Section 4.4.2.

Appendix 5.1 Mean standardized familiarity ratings for the Job ad stimuli

		Familiarity ratings					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
1	40 uur per week	0.99	(0.77)	0.97	(0.56)	1.13	(0.57)
2	voor meer informatie	0.70	(0.88)	0.79	(0.71)	0.94	(0.77)
3	kennis en ervaring	0.69	(0.79)	0.56	(0.85)	0.37	(0.75)
4	hoog in het vaandel	0.20	(0.69)	0.11	(0.79)	0.17	(0.86)
5	werving en selectie	1.19	(0.65)	0.90	(0.61)	-0.21	(0.85)
6	een vast dienstverband	0.99	(0.61)	0.52	(0.66)	-0.06	(0.62)
7	voor langere tijd	0.46	(0.75)	0.31	(0.67)	0.75	(0.54)
8	het eerste aanspreekpunt	0.41	(0.71)	0.23	(0.54)	-0.11	(0.70)
9	goede contactuele eigenschappen	0.58	(0.88)	0.20	(1.06)	-0.52	(0.84)
10	bij gebleken geschiktheid	0.28	(0.89)	-0.03	(0.73)	-0.60	(0.62)
11	academisch werk- en denkniveau	0.69	(0.58)	0.32	(0.65)	-0.32	(0.63)
12	een grote mate van zelfstandigheid	0.49	(0.52)	0.25	(0.68)	0.03	(0.74)

		Familiarity ratings					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
13	in een hecht team	0.62	(0.68)	0.47	(0.64)	0.66	(0.60)
14	een persoonlijk ontwikkelingsplan	0.34	(0.78)	0.13	(1.44)	-0.45	(0.72)
15	een sterk analytisch vermogen	0.88	(0.51)	0.75	(0.76)	0.17	(0.81)
16	met de mogelijkheid tot verlenging	0.53	(0.83)	0.26	(0.79)	-0.13	(0.61)
17	in de breedste zin van het woord	0.14	(0.82)	0.62	(1.00)	0.55	(0.77)
18	met een afstand tot de arbeidsmarkt	0.07	(0.60)	-0.82	(0.82)	-1.05	(0.54)
19	het geschetste profiel	0.33	(0.67)	0.12	(0.82)	-0.12	(0.65)
20	in de meest uiteenlopende sectoren	-0.38	(0.67)	-0.50	(0.67)	-0.84	(0.49)
21	een vliegende start	0.05	(0.91)	0.35	(1.06)	0.64	(0.83)
22	bewijs van goed gedrag	0.36	(0.71)	0.29	(0.75)	-0.03	(0.72)
23	conform de geldende CAO	0.33	(0.85)	0.03	(0.91)	-0.97	(0.68)
24	met behoud van uitkering	0.02	(0.78)	-0.31	(0.80)	-0.94	(0.48)

		Familarity ratings					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
25	bevoegd en bekwaam	-0.05	(0.81)	-0.23	(0.76)	-0.29	(0.73)
26	een integrale benadering	-0.70	(0.71)	-0.62	(1.00)	-1.33	(0.50)
27	naar aanleiding van de advertentie	0.28	(0.93)	0.28	(0.79)	0.09	(0.55)
28	eenvoudige administratieve werkzaamheden	0.17	(0.74)	-0.01	(0.78)	-0.13	(0.62)
29	een scherpe blik	0.03	(0.65)	0.56	(0.89)	0.62	(0.72)
30	buiten de geijkte paden	-0.34	(0.92)	-0.60	(1.07)	-1.14	(0.75)
31	affiniteit met het onderwerp	0.02	(0.73)	0.35	(0.67)	-0.29	(1.06)
32	een internationale speler van formaat	-0.83	(0.80)	-0.76	(0.89)	-0.44	(1.05)
33	een flinke portie lef	-0.49	(0.82)	-0.14	(0.72)	-0.26	(0.67)
34	met bewezen kwaliteiten	-0.08	(1.00)	-0.23	(0.94)	-0.48	(0.61)
35	een collegiale opstelling	-0.18	(0.78)	-0.30	(0.81)	-0.86	(0.68)

Appendix 5.2 Mean standardized familiarity ratings for the News report stimuli

		Familiarity ratings					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
36	de Tweede Kamer	0.59	(0.80)	0.74	(0.79)	1.23	(0.89)
37	wetenschap en techniek	-0.72	(1.03)	-0.50	(0.80)	-0.51	(0.65)
38	verkeer en vervoer	-0.63	(1.00)	-0.52	(0.91)	-0.46	(0.86)
39	in elk geval	0.67	(0.73)	0.83	(0.74)	1.44	(0.55)
40	in de Verenigde Staten	0.10	(0.84)	0.67	(0.67)	1.38	(0.78)
41	het openbaar ministerie	0.03	(0.95)	0.42	(1.10)	0.80	(0.66)
42	de negentiende eeuw	-0.15	(1.11)	0.18	(1.21)	1.04	(0.98)
43	de raad van bestuur	-0.02	(0.89)	0.37	(0.86)	0.32	(0.79)
44	aan de andere kant	0.53	(0.75)	0.67	(0.65)	1.07	(0.53)
45	evenementen en manifestaties	-1.17	(0.81)	-1.17	(0.86)	-1.00	(0.55)
46	het dagelijks leven	0.34	(0.77)	0.59	(0.59)	1.12	(0.61)
47	op een gegeven moment	0.35	(1.03)	0.66	(0.71)	1.37	(0.91)

	Cue	Familiarity ratings					
		Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
48	met terugwerkende kracht	0.25	(0.79)	0.41	(0.74)	0.33	(0.68)
49	in volle gang	0.03	(0.84)	0.38	(0.80)	0.77	(0.74)
50	een doorn in het oog	0.01	(0.94)	0.25	(1.03)	0.05	(0.96)
51	op geen enkele wijze	-0.26	(0.80)	-0.09	(0.71)	0.18	(0.67)
52	aan het begin van het seizoen	-0.17	(0.78)	-0.30	(0.72)	0.28	(0.60)
53	de lokale bevolking	0.06	(0.88)	0.01	(0.69)	0.46	(0.63)
54	het centrum van de stad	0.22	(0.86)	0.16	(0.90)	0.72	(0.70)
55	correcties en aanvullingen	-0.39	(0.73)	-0.26	(1.40)	-0.07	(0.64)
56	de opvang van asielzoekers	-0.12	(0.85)	-0.06	(0.69)	0.02	(0.68)
57	de traditionele partijen	-0.62	(1.63)	-0.63	(0.67)	-0.53	(0.74)
58	op last van de rechter	-0.96	(0.82)	-0.72	(0.71)	-1.00	(0.70)
59	in de huidige situatie	0.34	(0.75)	0.32	(0.48)	0.67	(0.57)

		Familarity ratings					
	Cue	Recruiters		Job-seekers		Inexperienced	
		M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
60	een onafhankelijke commissie	-0.53	(0.74)	-0.39	(0.65)	-0.76	(0.55)
61	een criminele afrekening	-0.93	(0.83)	-0.56	(0.84)	-0.42	(0.75)
62	de koninklijke loge	-1.39	(0.80)	-1.72	(0.87)	-1.31	(0.67)
63	een ingrijpende herstructurering	-0.78	(1.04)	-0.65	(0.69)	-0.72	(0.49)
64	op weg naar de top	-0.04	(0.82)	-0.14	(0.79)	0.29	(0.58)
65	in het belang van het kind	-0.43	(1.01)	-0.17	(0.80)	0.35	(0.77)
66	aan de vooravond van een revolutie	-1.21	(0.67)	-1.08	(0.87)	-1.38	(0.50)
67	de uitkomsten van het rapport	-0.10	(0.63)	-0.27	(0.72)	0.00	(0.62)
68	met hernieuwde energie	-0.52	(1.18)	-0.84	(1.35)	-0.36	(0.73)
69	een ongekende vrijheid	-0.44	(1.02)	-0.67	(0.87)	-0.16	(0.62)
70	een luxe jacht	-0.74	(0.73)	-0.72	(0.88)	0.24	(0.66)

Appendix 5.3 Linear mixed-effects models fitted to standardized familiarity ratings (Magnitude Estimation task)

We fitted linear mixed-effects models (Baayen et al. 2008), using the LMER function from the lme4 package in R (version 3.3.3; CRAN project; R Core Team, 2017), to the standardized familiarity ratings. We investigated to what extent these ratings can be predicted by corpus-based phrase frequency (LogFreqPhrase) and lemma frequency of the final word in the phrase (LogFreqLemma), whether or not a participant expected the final word to occur given the preceding words (TargetMentioned); and the time it took the participant to start pronouncing the target word when presented following the cue (VOT). In addition, we examined whether there are effects of PresentationOrder and Block. The fixed effects were standardized. We incorporated a random intercept for items to account for between-item variation. A by-participant random intercept was not included, because after the Z-score transformation all participants' scores have a mean of 0. We then added fixed effects one by one and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in familiarity ratings.

We started with LogFreqPhrase, which significantly contributed to the fit of the model ($\chi^2(1) = 35.14$, $p < .001$). We then added LogFreqLemma ($\chi^2(1) = 2.50$, $p = .11$). Given that it did not improve model fit, we left out this predictor. We proceeded with TargetMentioned ($\chi^2(1) = 283.00$, $p < .001$), followed by VOT ($\chi^2(1) = 8.90$, $p < .01$.), each of which was found to improve model fit. PresentationOrder did not contribute to the fit of the model ($\chi^2(1) = 3.77$, $p = .06$); Block did ($\chi^2(1) = 6.30$, $p < .05$). We then included the interaction terms LogFreqPhrase x TargetMentioned ($\chi^2(1) = 16.37$, $p < .001$), and VOT x TargetMentioned ($\chi^2(1) = 7.78$, $p < .01$). Finally, none of the two-way interactions of Block and the other predictors in the model was found to improve model fit (Block x LogFreqPhrase ($\chi^2(1) = 2.56$, $p = .11$); Block x TargetMentioned ($\chi^2(1) = 0.22$, $p = .64$); Block x VOT ($\chi^2(1) = 0.25$, $p = .62$)).

The model selection procedure thus resulted in a model comprising LogFreqPhrase, TargetMentioned, VOT, Block, LogFreqPhrase x TargetMentioned, and VOT x TargetMentioned. For all of these fixed effects, we included a by-participant random slope. For the factor VOT we also added a by-item random slope. There are no other by-item random slopes, because each item has only one phrase frequency, and occurred in only one of the two blocks. Furthermore, there are items no one had mentioned in the completion task, thus prohibiting by-item random slopes for TargetMentioned. Within these limits, a model with a full random effect structure was constructed following Barr et al. (2013). Subsequently, we excluded random slopes with the lowest variance step by step until a further reduction would imply a significant loss in the goodness of fit of the model (Matuschek et al. 2017). Model comparisons indicated that the inclusion of the by-participant random slopes for LogFreqPhrase, TargetMentioned, and Block was justified by the data (χ2(4) = 97.25, *p* = .001). The variance explained by this model is 37% (R^2m = .18, R^2c = .37).[36] Confidence intervals were estimated via parametric bootstrapping over 10000 iterations (Bates et al. 2015). We first ran the model using *Target not mentioned* as the reference condition and then *Target mentioned*. The outcomes are presented in Tables 5.1 and 5.2 in Section 5.4.

[36] R^2m (marginal R^2 coefficient) represents the amount of variance explained by the fixed effects; R^2c (conditional R^2 coefficient) is interpreted as variance explained by both fixed and random effects (i.e. the full model) (Johnson 2014).

Summary

This dissertation presents research into variation between and within participants in their metalinguistic judgments about, and processing of, multi-word sequences. Numerous studies provide evidence that language users are sensitive to the likelihood of words to co-occur and that they make use of this information in language acquisition and processing (for overviews see Diessel 2007; Gries & Divjak 2012; Jurafsky et al. 2001; Kuperberg & Jaeger 2016). The more frequently a string of words is used, the more quickly and easily the sequence is retrieved and processed and the more familiar it is considered to be. This suggests that usage frequency affects our mental representations of language: more experience with a linguistic construction makes it more strongly entrenched in the speaker's mental lexicon, which in turn influences the probability that the construction will be used, the speed with which it is processed, and the speaker's metalinguistic knowledge regarding its use.

If usage-based models of linguistic representations (Barlow & Kemmer 2000; Bybee 2006; Goldberg 2006; Langacker 1987; Schmid 2007; Tomasello 2003) are correct in positing such a strong link between usage frequency and entrenchment, it follows that the extent to which a linguistic construction is entrenched varies from person to person, as well as over time. That is, since language users differ in their linguistic experiences, there are likely to be differences in entrenchment across individuals. Furthermore, a language user gains new linguistic experiences over time, and usage-based linguistics predicts mental representations of language to change accordingly. There is a shortage of empirical data on these types of variation, though. As I discuss in more detail in Chapter 1, the past five decades have seen a wealth of studies yielding evidence in support of usage-based theories of language acquisition and processing, but these studies have paid little attention to inter- and intra-individual variation in adult native speakers. A central aim of the studies presented in this dissertation is to show that insight into these types of variation is a prerequisite for a veridical description of mental representations of language.

Chapters 2 and 3 present two studies that examine inter- and intra-individual variation in metalinguistic judgments. The latter was investigated by means of a test-retest design: participants performed the same task twice within the space of one to three weeks. In both studies, native speakers of Dutch were asked to assign familiarity ratings to a set of prepositional phrases that cover a wide range of corpus frequencies (e.g. *op de bank* 'on the couch / in the bank', *in de lucht* 'in the air'). In the study reported on in Chapter 2, 44 phrases were presented in

isolation as well as in a sentential context, to investigate whether context affects perceived degree of familiarity and inter- and intra-individual variation in judgments. The participants assigned ratings using the method of Magnitude Estimation (Bard et al. 1996). Aggregated scores (averaged over 86 participants) are remarkably consistent (Pearson's *r* = .97), and there is a significant relationship between familiarity ratings and corpus frequencies of the phrases. At the same time, there is considerable variation between and within participants. Context does not reduce this variation. As random noise does not seem to account for the patterns of variation in the data, I propose to consider the possibility that intra-individual variation is a genuine property of one's metalinguistic representations and ultimately one's linguistic representations. This implies that the difference between people's ratings at one point in time cannot be interpreted straightforwardly as *the* difference in their linguistic representations. A more complete and more faithful impression requires multiple measurements.

Chapter 3 starts by describing how, in various fields of linguistics, variation has been overlooked, looked at from a limited perspective (e.g. variation being simply the result of irrelevant performance factors), or considered troublesome. I then argue that it is both feasible and valuable to study different types of variation. To illustrate this, I conducted an experiment in which 91 participants assigned familiarity ratings to 79 prepositional phrases. They performed the task twice within a couple of weeks, using either a 7-point Likert scale or a Magnitude Estimation scale. The research design employed here thus yielded data on variation across items, across participants, across time, and across rating methods. I explicate the principles according to which the different types of variation can be considered information about mental representation, and I show how they can be used to test hypotheses regarding linguistic representations.

The results indicate that familiarity judgments form methodologically reliable, useful data in linguistic research. The ratings obtained with one scale were corroborated by the ratings on the other scale. In addition, there was a near perfect Time1–Time2 correlation of the mean ratings in all experimental conditions, and in all conditions the majority of the participants had high self-correlation scores. Furthermore, the data show a clear correlation between familiarity ratings and corpus frequencies.

Similar to the dataset analyzed in Chapter 2, the familiarity ratings display inter- and intra-individual variation. Usage-based exemplar models (Goldinger 1996; Hintzman 1986; Pierrehumbert 2001) naturally accommodate such variation. In these models, linguistic representations consist of a continually updating set of exemplars. An exemplar is not a tape recording stored in memory, but a multidimensional, detail-rich representation that follows from a process of analysis and categorization (Taylor 2012). While the judgment task requires people to indicate the position of a given item on a scale of familiarity by means of a single value, its familiarity for a particular speaker may best be viewed as a moving target located in a region that may be narrower or wider. In that case, there is not just one true value, but a range of scores that constitute true expressions of an item's familiarity. Variation in judgment across time is not noise then, but a reflection of the dynamic character of cognitive representations as more, or less, densely populated clouds of exemplars that vary in strength depending on frequency and recency of use.

Chapter 3 concludes with a discussion of the similarities and differences between Magnitude Estimation (ME) and Likert scale ratings. In several respects, the two scales yielded similar outcomes, but there are also differences that ought to be taken into account when selecting a particular scale. Likert ratings, unlike ME ratings, make it possible to determine whether participants consider the majority of items to be familiar (or unfamiliar), and whether they consider the entire set of stimuli more familiar the second time (as a result of the exposure in the test sessions, for example). A disadvantage of using a Likert scale is the risk that the number of response options does not match the degrees of familiarity as perceived by the participants, which could result in a loss of information. ME allows participants to distinguish as many degrees as they feel relevant. When using ME, the vast majority of the participants in the study reported here (83.3%) distinguished more than seven degrees, indicating that a 7-point Likert scale may not be optimal for the construct and the set of stimuli used here.

In **Chapter 4 and 5**, I examine inter- and intra-individual variation by means of three experiments that I conducted with three groups of participants: 40 recruiters, 40 job-seekers, and 42 people not (yet) looking for a job (henceforth referred to as Inexperienced). These groups can be expected to differ in experience with word sequences that typically occur in job ads (e.g. *goede contactuele eigenschappen* 'good communication skills'); they are not expected to differ systematically in experience with word sequences characteristic of news reports (e.g. *de Tweede Kamer* 'the House of Representatives'). The word sequences were used as stimuli in a completion task, a voice onset time (VOT) experiment, and a familiarity judgment task. I thus examined the relationship between amount of experience with a particular register and (i) the expectations people generate about upcoming words when faced with word strings characteristic of that register; (ii) the speed with which they process such word strings; and (iii) how familiar they consider these word strings to be. Furthermore, I investigated the relationships between data elicited from an individual participant in different types of psycholinguistic tasks using the same stimuli. More specifically, I compared participant-based measures, on the one hand, and measures based on amalgamated data of different people, on the other, as predictors of performance in psycholinguistic tasks. This provides insight into individual variation and the merits of going beyond amalgamated data.

Chapter 4 reports on the completion task and the VOT task. In the completion task, the participants were shown incomplete phrases (e.g. *goede contactuele …* 'good communication …') and for each stimulus they listed all complements that came to mind within five seconds. This task yielded information on the expectations people generate about upcoming words. Their responses were compared with the complements observed in a job ad corpus and the Twente News Corpus. The analyses revealed that on the News Report items, the groups did not differ significantly from each other in the proportion of responses that correspond to a complement observed in the Twente News Corpus. On the Job ad stimuli, by contrast, the groups did differ significantly, as hypothesized. The Recruiters' responses corresponded significantly more often to complements observed in the Job ad corpus than the Job-seekers' responses. The Job-seekers' responses, in turn, corresponded significantly more often to a complement in the Job ad corpus than the responses of the Inexperienced participants. The results indicate that there are differences in participants' knowledge of multi-word units which are related to their degree of experience with these word sequences.

In the subsequent VOT experiment, the participants were presented with the same cues (e.g. *goede contactuele …* 'good communication …'), followed by a specific target word (e.g. *eigenschappen* 'skills'), which they had to read aloud as quickly as possible. The voice onset times indicate how much time it takes to process the target word in the given context. According to prediction-based processing models (Bar 2007; A. Clark 2013; Huettig 2015; Kuperberg & Jaeger 2016; Kutas et al. 2011), the target will be easier to recognize and process when it consists of a word that the participant expected than when it consists of an unexpected word. Most studies to date quantify a word's predictability by means of cloze probabilities and surprisal estimates, which are based on amalgamations of data of various speakers and thus disregard variation across speakers. Having participants perform both a completion task and a VOT task made it possible to relate reaction times to participants' own expectations.

Firstly, the analyses revealed that the majority of the Recruiters and the Job-seekers responded faster to the Job ad items than to the News report items, while it was exactly the other way around for the vast majority of the Inexperienced participants. I then examined to what extent variation in VOTs across items and across participants could be explained by different measures of word predictability, while accounting for characteristics of the target words (i.e. word length and word frequency) and the experimental design (i.e. presentation order and block). Whether or not participants had mentioned the target significantly affected voice onset times. What is more, this predictive pre-activation, as captured by the variable TargetMentioned, was found to facilitate processing to such an extent that word frequency could not exert any additional accelerating influence. This demonstrates the impact of context-sensitive prediction on subsequent processing. Perhaps even more interesting is that the variable TargetMentioned had an effect on voice onset times over and above the effect of ClozeProbability. This shows the added value of going beyond amalgamated data. While this may not come across as surprising, it is seldomly shown or exploited in research on prediction-based processing.

After having completed the VOT task, the participants assigned familiarity ratings to the word sequences using Magnitude Estimation. In Chapter 5, I analyze the judgment data in relation to the data from the completion task and the VOT task as well as corpus frequencies. In this way, I examine whether the degree to which linguistic constructions are entrenched in the participants' minds manifests itself not just in processing but also in metalinguistic judgments. In other words, are these degrees of entrenchment part of one's explicit knowledge and can metalinguistic judgments be used to gain insight into entrenchment? On the one hand, "judgments are the results of linguistic and cognitive processes, by which people attempt to process sentences and then make metalinguistic judgments on the results of those acts of processing (…) Thus, they implicate the same linguistic representations involved in all acts of processing", as Branigan and Pickering (2017: 4) contend. On the other hand, judgments may be influenced by knowledge and beliefs (Dąbrowska 2016a) and reflect decision-making biases (Branigan & Pickering 2017) which are not involved in language processing. Various researchers are concerned that introspections cannot yield accurate insights into subconscious cognitive processes (e.g. Gibbs 2006; Roehr 2008; Stubbs 1993).

Prior research has examined the relationship between familiarity ratings and various kinds of psycholinguistic data. A limitation of those studies is that the sets of familiarity ratings come from different people than the datasets indicating performance in processing tasks. Consequently, we cannot tell whether a discrepancy between familiarity judgments and processing data reflects the fact that different tasks tap into different processes and knowledge, or whether it reflects individual variation in linguistic representations. By having participants perform both a judgment task and processing tasks, I was able to differentiate between the two.

Firstly, the results show that differences in experiences with a particular register were reflected in the familiarity ratings that participants assigned to phrases characteristic of that register. The vast majority of the Recruiters considered the Job ad phrases to be more familiar than the News report phrases, while for the Inexperienced participants it was the other way around. Secondly, individual participants' data from the completion task and the VOT task are significant predictors of the familiarity ratings they assigned to the stimuli. This indicates that familiarity judgments and other types of psycholinguistic data tap into the same mental representations of language, and that familiarity ratings form useful data to gain insight into these representations.

The dissertation concludes with two chapters in which I reflect on the studies I conducted. In Chapter 6, I focus on the methodological lessons that can be learned from them. The chapter highlights the merits of multi-method research in linguistics and offers an overview of key considerations in the design of such research. It discusses methodological and practical concerns in the selection of corpus data, metrics to analyze corpus data, stimuli, experimental tasks, and participants.

In Chapter 7, I consider the theoretical implications of my findings. The results indicate that there are systematic differences in participants' knowledge and processing of multi-word units which are related to their degree of experience with these word sequences. This forms empirical support for hypotheses that follows from usage-based theories of linguistic knowledge and language processing. Furthermore, an individual's performance in one experiment was shown to be a significant predictor of performance in another experiment, on top of measures based on amalgamated data of different people (i.e. corpus-based frequencies, surprisal, cloze probabilities). In other words, participant-based measures proved to have unique additional explanatory power. This demonstrates the existence of systematic, measurable inter-individual variation in behavioral indices of cognitive routinization. Variation is ubiquitous, but, crucially, not random. One of the important tasks that we face when we want to arrive at accurate theories of linguistic representation and processing is to define the factors that determine the degrees of variation between individuals, and this requires going beyond amalgamated data.

In addition to inter-individual variation, there is evidence of intra-individual variation which, too, points to the dynamic character of mental representations of language. Most psycholinguistic tasks that try to tap into the degree of entrenchment of a linguistic unit in the mind of a speaker, express this in a single value (e.g. a rating, a reaction time). However, if cognitive representations can best be viewed as more, or less, densely populated clouds of exemplars that vary in strength depending on frequency and recency of use, a single score yields an incomplete picture. Therefore, I not only advocate attending to variation across participants, I also urge cognitive linguists to carry out multiple measurements per participant.

To conclude, I sketch three compelling directions for future research that build on the work presented in this dissertation. I propose, first of all, to further develop participant-based measures. In my studies, I converted the completion task responses into a variable that indicates for each participant individually whether the target word had been mentioned or not (TargetMentioned). It proved to be a valuable measure. However, as a binary variable, it does not account for gradient differences in the degree to which words are expected to occur. I provide suggestions as to how the potential of participant-based data can be explored.

Secondly, I propose to follow participants in the course of a few weeks or months, extending the test-retest design. This can provide additional insights into the effects of usage frequency on processing speed and perceived degree of familiarity. It is clear by now that frequency is a key factor. What is not so clear, is to what extent recency of use matters; whether it makes a difference whether you used a linguistic item once or twice that day; and whether this works differently for low-frequency items compared to high-frequency ones.

Thirdly, I propose to examine (partially) schematic constructions in addition to lexically specific ones. On a usage-based account, mental representations of (partially) schematic constructions are dynamic in nature too, just like representations of lexically specific constructions such as morphemes, complex words, and multi-word units. All representations are taken to emerge from, and are continuously shaped by, experience with language together with general cognitive skills such as categorization, schematization, and chunking. However, schematic constructions tend to have a more general meaning, a wider range of usage contexts, and a higher frequency of occurrence than lexically specific constructions, which may result in less inter- and intra-individual variability. What should also be taken into account, is that speakers may differ in cognitive abilities, such as language analytic ability, statistical learning ability, fluid intelligence, and cognitive motivation (Dąbrowska 2018; Misyak & Christiansen 2012). Both linguistic experiences and cognitive abilities appear to influence the process of schematization and speakers' knowledge of grammatical constructions. There are indications that this does not hold for collocational knowledge in the same way (Dąbrowska 2018). While representations of words, multi-word units, and grammatical patterns can still be construed as constructions that emerge from linguistic experience together with general cognitive skills, they may differ in the extent to which they rely on various cognitive and experiential factors. Research that aims to advance our understanding of the contributions of these factors must pay attention to individual differences. I hope this dissertation contributes to this research agenda by demonstrating that it is feasible and valuable to attend to inter- and intra-individual variation and by sparking linguists’ enthusiasm for such an approach.

Samenvatting

Stel, aan een groep mensen wordt de zin *Bij gelijke geschiktheid gaat onze voorkeur uit naar een vrouwelijke kandidaat* voorgelegd. In hoeverre verschillen zij van elkaar in de manier waarop ze deze zin verwerken, en kunnen we deze verschillen verklaren? Lange tijd beschouwden taalkundigen woorden en grammaticale regels als de bouwstenen in taal. In de afgelopen vijftig jaar is echter duidelijk geworden dat dat niet volstaat als beschrijving van de mentale organisatie van taal. We beschikken over een veel gevarieerdere set aan talige eenheden. Een zin als *Bij gelijke geschiktheid gaat onze voorkeur uit naar een vrouwelijke kandidaat* kan geproduceerd en begrepen worden door de losse woorden en de syntactische structuur waarin ze zijn ingebed te activeren, maar taalgebruikers kunnen ook grotere eenheden gebruiken. Ze kunnen bijvoorbeeld gebruik maken van woordcombinaties (*multi-word units* zoals *bij gelijke geschiktheid*) en gedeeltelijk schematische eenheden (zoals *gaat* LIDWOORD/BEZITTELIJK VNW *voorkeur uit naar* NAAMWOORDGROEP). Psycholinguïstisch onderzoek heeft aangetoond dat sommige van dergelijke constructies sneller worden verwerkt, gemakkelijker worden herinnerd, en vertrouwder aandoen dan andere. Dit suggereert dat ze verschillen in de mate waarin ze verankerd zijn in onze taalkennis – met andere woorden, de mate van *entrenchment* varieert. Gebruiksfrequentie lijkt een sleutelrol te spelen in het proces van entrenchment: hoe vaker een talige constructie gebruikt wordt, hoe sterker deze verankerd wordt in het mentale lexicon van de taalgebruiker, waardoor het makkelijker wordt om de constructie te activeren en te verwerken.

Gebruiksgebaseerde modellen van mentale representaties van taal (Barlow & Kemmer 2000; Bybee 2006; Goldberg 2006; Langacker 1987; Schmid 2007; Tomasello 2003) stellen dat er een sterk verband is tussen gebruiksfrequentie en entrenchment. Als dit werkelijk zo is, dan varieert de mate waarin in een constructie verankerd is zowel van persoon tot persoon, als in de loop der tijd. Variatie in entrenchment tussen mensen komt voort uit het feit dat taalgebruikers van elkaar verschillen in de frequentie waarmee ze bepaalde constructies tegenkomen en gebruiken. Variatie door de tijd heen volgt uit het feit dat taalgebruikers nieuwe ervaringen met taal opdoen gedurende hun leven. Volgens gebruiksgebaseerde modellen veranderen mentale representaties van taal mee: toenemend gebruik leidt tot sterkere verankering; de representatie verzwakt als een constructie een tijd lang niet gebruikt wordt (Langacker 1987: 59). Empirische data over deze vormen van variatie zijn echter schaars. In Hoofdstuk 1 beschrijf ik dat er in de laatste vijf decennia veel onderzoek heeft plaatsgevonden waarvan de uitkomsten in lijn zijn met gebruiksgebaseerde theorieën over taalverwerving

en –verwerking. Het merendeel van deze studies heeft echter weinig aandacht besteed aan variatie tussen en binnen volwassen moedertaalsprekers. Het doel van de studies in dit proefschrift is aan te tonen dat inzicht in deze typen variatie noodzakelijk is om te komen tot een waarheidsgetrouwe beschrijving van mentale representaties van taal. Ik doe dit door de variatie tussen en binnen participanten in metalinguïstische oordelen over, en verwerking van meerwoordsconstructies te onderzoeken.

Hoofdstukken 2 en 3 rapporteren over twee studies naar inter- en intra-individuele variatie in metalinguïstische oordelen (oordelen waarbij je reflecteert op taal, taalgebruik, en taalkennis). Door de oordelentaak bij verschillende mensen af te nemen is informatie verkregen over interindividuele variatie. Intra-individuele variatie is onderzocht door deelnemers dezelfde taak twee keer te laten uitvoeren in een periode van één tot drie weken. In beide studies hebben moedertaalsprekers van het Nederlands vertrouwdheidsoordelen toegekend aan voorzetselgroepen (bijv. *op de bank*, *in de lucht*). Deze woordcombinaties varieerden in de frequentie waarmee ze voorkomen in een groot corpus van hedendaags Nederlands taalgebruik. In de studie die beschreven wordt in Hoofdstuk 2 zijn 44 voorzetselgroepen gepresenteerd als losse woordcombinaties en tevens ingebed in een zin, om na te gaan of context van invloed is op het gevoel van vertrouwdheid en op de variatie in oordelen. De participanten kenden scores toe aan de hand van een methode die *Magnitude Estimation* heet (Bard et al. 1996). De geaggregeerde waardes, waarbij het gemiddelde werd genomen van de scores van 86 participanten, bleken opmerkelijk consistent (Pearson's *r* = .97), en er was een significant verband tussen de vertrouwdheidsscores en corpusfrequenties (hogere frequenties gaan gepaard met hogere scores). Tegelijkertijd was er sprake van aanzienlijke variatie tussen en binnen participanten in oordelen. Het toevoegen van een zinscontext verminderde deze variatie niet. Er zijn taalkundigen (bijv. Featherston 2007) die van mening zijn dat inter- en intra-individuele variatie in metalinguïstische oordelen ruis is, die er uit gefilterd kan worden door met geaggregeerde scores te werken. De variatie in mijn dataset vertoonde echter patronen die niet verklaard lijken te kunnen worden in termen van willekeurige ruis. Daarom stel ik voor om de mogelijkheid te overwegen dat intra-individuele variatie een echt kenmerk is van metalinguïstische representaties en zelfs van alle soorten talige representaties. Variatie in oordelen van moment tot moment zou een reflectie kunnen zijn van de dynamiek van talige representaties. Dit impliceert dat het verschil tussen de oordelen van twee mensen op één bepaald moment niet zomaar beschouwd kan worden als hét verschil tussen hun mentale representaties van taal. Op een ander moment kan het plaatje er namelijk anders uitzien. Voor een vollediger en waarheidsgetrouwer beeld zijn meerdere metingen nodig.

In **Hoofstuk 3** beschrijf ik hoe, in verscheidene gebieden binnen de taalkunde, variatie over het hoofd werd gezien, beschouwd werd als simpelweg het gevolg van irrelevante factoren (zoals beperkingen van het werkgeheugen en vergissingen), of als lastig werd ervaren. Vervolgens bepleit ik dat het mogelijk en waardevol is om verschillende typen variatie te bestuderen. Dit illustreer ik aan de hand van een experiment waarbij 91 deelnemers 79 voorzetselgroepen beoordeelden op vertrouwdheid. Ze voerden deze taak tweemaal uit, waarbij ze gebruik maakten van ofwel een 7-puntslikertschaal, ofwel een Magnitude Estimation schaal. Zo werden gegevens verkregen over variatie tussen items (de voorzetselgroepen in dit geval), tussen participanten, tussen meetmomenten, en tussen meetmethodes (Likert vs. Magnitude Estimation). Ik zet uiteen hoe de verschillende typen variatie informatie kunnen verschaffen over mentale representaties van taal, en ik toon hoe ze gebruikt kunnen worden om hypotheses over representaties te toetsen.

De uitkomsten van deze studie geven aan dat vertrouwdheidsoordelen methodologisch betrouwbare, bruikbare data zijn in taalkundig onderzoek. De scores die met de ene schaal verkregen waren, werden bevestigd door de scores op de andere schaal. Er was bovendien in alle experimentele condities een vrijwel perfecte correlatie tussen de gemiddelde scores op moment 1 en moment 2. Daarnaast had, in iedere conditie, de meerderheid van de participanten hoge zelf-correlatie waarden (m.a.w. iemands oordelen op moment 2 correleerden sterk met diens eigen oordelen op moment 1). Ook was er sprake van een duidelijk verband tussen vertrouwdheidsoordelen en corpusfrequenties.

De oordelen vertoonden, net als de dataset in Hoofdstuk 2, inter- en intra-individuele variatie. Gebruiksgebaseerde *exemplar* modellen (Goldinger 1996; Hintzman 1986; Pierrehumbert 2001) bieden van nature ruimte voor dergelijke variatie. In deze modellen bestaan mentale representaties van taal uit een set *exemplars* die continu geüpdatet wordt. Een *exemplar* is niet een kleine bandopname die opgeslagen wordt in je geheugen, maar een multidimensionale, detailrijke representatie die volgt uit een proces van analyse en categorisatie (Taylor 2012). In de oordelentaak moeten deelnemers de positie van een item op een schaal van vertrouwdheid uitdrukken in één getal, terwijl de vertrouwdheid misschien eerder een bewegend doel is in een ruimte die meer of minder breed kan zijn. In dat geval is er niet slechts één ware score, maar een reeks waarden die de vertrouwdheid van een item uitdrukken. Variatie in scores van moment tot moment hoeft geen ruis te zijn; het kan de weerslag zijn van het dynamische karakter van cognitieve representaties als meer, of minder, compacte clusters van *exemplars* die variëren in sterkte afhankelijk van hoe frequent en hoe recent bepaalde constructies zijn gebruikt.

Hoofdstuk 3 besluit met een bespreking van de overeenkomsten en verschillen tussen oordelen die met behulp van Magnitude Estimation (ME) uitgedrukt worden en Likertschaaloordelen. In verscheidene opzichten leverden de twee schalen vergelijkbare uitkomsten op, maar er zijn ook verschillen waar rekening mee moet worden houden bij het kiezen van een schaal. Zo kan alleen met de Likertschaaloordelen bepaald worden of respondenten de meerderheid van de items als vertrouwd (of niet vertrouwd) beschouwen, en of zij de gehele set items de tweede keer vertrouwder achten (bijv. door het bezig zijn met de items tijdens de experimenten). Een nadeel van het gebruiken van een Likertschaal is het risico dat het aantal responseopties niet overeenkomt met de vertrouwdheidsgradaties die de participanten reëel achten, waardoor er informatie verloren kan gaan. ME staat participanten toe om precies het aantal gradaties te onderscheiden dat zij relevant vinden. In het onderzoek dat beschreven wordt in Hoofdstuk 3, onderscheidde het overgrote deel (83.3%) van de deelnemers meer dan zeven gradaties bij het gebruik van ME, wat er op wijst dat een 7-puntslikertschaal wellicht niet optimaal is voor het construct (vertrouwdheidsoordelen) en de items (de 79 voorzetselgroepen) die hier gebruikt werden.

In **Hoofdstukken 4 en 5** onderzoek ik inter- en intra-individuele variatie door middel van drie experimenten die ik heb afgenomen bij drie groepen deelnemers: 40 recruiters en HR-managers, 40 werkzoekenden, en 42 studenten die zelden of nooit vacatureteksten hadden gelezen (hierna de onervaren deelnemers genoemd). Het is aannemelijk dat deze groepen verschillen in ervaring met woordcombinaties die typisch zijn voor vacatureteksten (bijv. *goede contactuele eigenschappen, werving en selectie*); er worden geen systematische verschillen verwacht tussen de groepen in ervaring met woordcombinaties die kenmerkend zijn voor nieuwsberichten (bijv. *de Tweede Kamer*, *correcties en aanvullingen*). De woordcombinaties werden gebruikt als stimuli in een aanvultaak, een *voice onset time* (VOT) experiment, en een vertrouwdheidsoordelentaak. Aldus onderzocht ik of er een verband is tussen enerzijds de mate van ervaring met een bepaald register en anderzijds (i) de verwachtingen die mensen genereren over woorden die mogelijk volgen wanneer ze woordsequenties zien die kenmerkend zijn voor dat register; (ii) de snelheid waarmee ze dergelijke woordcombinaties verwerken; en (iii) hoe vertrouwd deze woordcombinaties voor hen zijn. Ook onderzocht ik hoe verschillende soorten data van één participant, verkregen in verschillende psycholinguïstische taken, zich tot elkaar verhouden. Ik heb maten die gebaseerd zijn op data van een individuele participant vergeleken met maten die gebaseerd zijn op data van verschillende mensen. Dit verschaft inzicht in individuele variatie en de toegevoegde waarde van gepersonaliseerde maten ten opzichte van geaggregeerde data.

Hoofdstuk 4 doet verslag van de aanvultaak en de VOT-taak. In de aanvultaak kregen de deelnemers incomplete frases te zien (bijv. *goede contactuele …*). Bij iedere stimulus somden ze de aanvullingen op die binnen vijf seconden in hen opkwamen. Deze taak levert informatie op over de verwachtingen die iemand genereert over woorden die kunnen volgen. De antwoorden werden vergeleken met de aanvullingen die voorkomen in een corpus met vacatureteksten en het Twente Nieuws Corpus. De analyses wezen uit dat er wat betreft de nieuwsberichtstimuli geen significante verschillen waren tussen de groepen in de proportie van antwoorden die corresponderen met een aanvulling in het Twente Nieuws Corpus. Op de vacaturestimuli, daarentegen, waren er significante verschillen tussen de groepen, zoals verwacht. De responses van de recruiters kwamen significant vaker overeen met aanvullingen in het vacaturecorpus dan de responses van de werkzoekenden. De responses van de werkzoekenden kwamen op hun beurt weer significant vaker overeen met aanvullingen in het vacaturecorpus dan de responses van de onervaren deelnemers. Deze bevindingen tonen aan dat er verschillen zijn tussen de participanten in kennis van meerwoordsconstructies, en dat die verschillen samenhangen met de mate waarin zij ervaring hebben met deze constructies.

In de daaropvolgende VOT-taak kregen de participanten dezelfde woordsequenties te zien (bijv. *goede contactuele …*), dit keer gevolgd door een specifiek woord (bijv. *eigenschappen*) dat ze zo snel mogelijk moesten voorlezen. Ik berekende hoeveel milliseconden het duurde voor iemand het woord begon uit te spreken. Deze *voice onset time* geeft aan hoeveel tijd het kost om het woord te verwerken in de gegeven context. De hypothese is dat het woord gemakkelijker herkend en verwerkt kan worden als het reeds verwacht werd gegeven de context (*prediction-based processing models*, Bar 2007; A. Clark 2013; Huettig 2015; Kuperberg & Jaeger 2016; Kutas et al. 2011). In eerder onderzoek is de voorspelbaarheid van een woord gekwantificeerd door middel van *cloze probabilities* (het percentage van de deelnemers dat dat woord invulde in de gegeven context) en *surprisal estimates* (de mate waarin het woord afwijkt van de voorspellingen gegenereerd door taalmodellen die getraind zijn op corpus data). Deze maten zijn gebaseerd op data van een grote groep taalgebruikers en gaan dus voorbij aan inter-individuele variatie. Doordat iedere deelnemer aan mijn onderzoek zowel de aanvultaak als de VOT-taak maakte, kon ik het verband tussen reactietijden en iemands eigen verwachtingen onderzoeken.

Uit de analyses bleek dat de meerderheid van de recruiters en de werkzoekenden sneller reageerde op de vacature-items dan op de nieuwsberichtitems, terwijl het omgekeerde het geval was voor het overgrote deel van de onervaren participanten. Vervolgens heb ik onderzocht in hoeverre de variatie in reactietijden tussen items en tussen participanten verklaard kan worden door verschillende maten van de voorspelbaarheid van een woord, waarbij ik rekening hield met kenmerken van de woorden (woordlengte en woordfrequentie) en het onderzoeksontwerp (de volgorde waarin items gepresenteerd werden). De reactietijden in de VOT-taak bleken significant sneller te zijn als participanten het woord genoemd hadden tijdens de aanvultaak – dit laatste werd uitgedrukt in de variabele TargetMentioned. De pre-activatie van woorden tijdens het genereren van verwachtingen bleek de verwerking van de woorden in de VOT-taak zozeer te vergemakkelijken dat woordfrequentie hier niets meer aan toevoegde. Doorgaans worden hoogfrequente woorden sneller herkend en verwerkt dan laagfrequente woorden, maar als het woord reeds genoemd was tijdens de aanvultaak had woordfrequentie geen effect meer. Wellicht nog interessanter is dat de variabele TargetMentioned van invloed was op reactietijden bovenop het effect van ClozeProbability. Dit illustreert de toegevoegde waarde van maten die rekening houden met variatie tussen participanten.

Na de VOT-taak kenden de deelnemers aan de hand van *Magnitude Estimation* vertrouwdheidsscores toe aan de woordcombinaties. Hoofdstuk 5 beschrijft de analyse van de vertrouwdheidsoordelen in relatie tot de data uit de aanvultaak en de VOT-taak, en corpusfrequenties. Ik heb onderzocht of de mate waarin woordcombinaties verankerd zijn in de mentale representaties van de participanten niet alleen tot uitdrukking komt in de wijze waarop zij de woordcombinaties verwerken, maar ook in hun metalinguïstische oordelen. Is de mate van entrenchment onderdeel van iemands expliciete kennis en kunnen metalinguïstische oordelen inzicht verschaffen in entrenchment? Aan de ene kant zijn dergelijke oordelen het resultaat van cognitieve processen waarmee de taalinput verwerkt wordt en waarmee er gereflecteerd wordt op de uitkomsten van die verwerking. De oordelen doen daarmee een beroep op representaties van taal die ook in andere vormen van verwerking een rol spelen (Branigan & Pickering 2017: 4). Aan de andere kant zouden oordelen beïnvloed kunnen worden door kennis, overtuigingen, en biases die niet meespelen in taalverwerking (Dąbrowska 2016a; Branigan & Pickering 2017). Verscheidene onderzoekers zijn bezorgd dat introspectie geen accuraat inzicht kan verschaffen in onderbewuste cognitieve processen (o.a. Gibbs 2006; Roehr 2008; Stubbs 1993).

Er is al eerder onderzoek gedaan naar de relatie tussen vertrouwdheidsoordelen en verscheidene soorten psycholinguïstische data. Een beperking van die studies is dat de vertrouwdheidsoordelen van één groep mensen komen en de taalverwerkingsdata van een andere groep. Een discrepantie tussen oordelen en verwerkingsdata zou kunnen betekenen dat de taken een beroep doen op verschillende processen en kennis; het zou echter ook het gevolg kunnen zijn van individuele variatie in cognitieve representaties van taal. Aangezien in mijn onderzoek de verschillende soorten data afkomstig zijn van dezelfde participanten, kan ik een onderscheid maken tussen variatie tussen taken enerzijds en variatie tussen participanten anderzijds.

De verschillen tussen de groepen deelnemers in ervaring met een bepaald register bleken tot uitdrukking te komen in de vertrouwdheidsscores die ze toekenden aan woordcombinaties die kenmerkend zijn voor dat register. De overgrote meerderheid van de recruiters beschouwde de vacature-items namelijk als vertrouwder dan de nieuwsbericht-items, terwijl het omgekeerde het geval was voor de onervaren participanten. Uit de analyses bleek vervolgens dat Iemands eigen data uit de aanvultaak en de VOT-taak significante voorspellers waren voor de vertrouwdheidsoordelen die diegene toekende. Dit wijst erop dat vertrouwdheidsoordelen en andere soorten psycholinguïstische data een beroep doen op dezelfde mentale representaties van taal en dat vertrouwdheidsoordelen bruikbare data vormen om inzicht te verkrijgen in deze representaties.

In de laatste twee hoofdstukken reflecteer ik op de onderzoeken die ik heb uitgevoerd. In Hoofdstuk 6 ligt de focus op de methodologische lessen die getrokken kunnen worden uit mijn studies. Ik belicht de verdiensten van onderzoek waarin verscheidene methodes gecombineerd worden en ik bied een overzicht van de belangrijkste overwegingen in het ontwerp van dergelijk onderzoek. Aan bod komen methodologische en praktische kwesties met betrekking tot het selecteren van: corpusdata, metrieken om corpusdata te analyseren, stimuli, experimentele taken, en participanten.

In Hoofdstuk 7 ga ik in op de theoretische implicaties van mijn bevindingen. De resultaten geven aan dat er systematische verschillen zijn tussen mensen in kennis en verwerking van woordcombinaties, en dat die verschillen in verband staan met de mate van ervaring met deze woordcombinaties. Dit vormt empirische ondersteuning voor hypotheses die volgen uit gebruiksgebaseerde theorieën over taalkennis en –verwerking. Voorts bleken de data van een participant afkomstig uit één type experiment een significante voorspeller te zijn voor diens prestaties in volgende experimenten, bovenop maten die gebaseerd zijn op data van een grote groep taalgebruikers (corpusfrequenties, *surprisal estimates, cloze probabilities*). Met andere woorden, gepersonaliseerde maten hebben unieke verklarende kracht. Dit toont aan dat er sprake is van systematische, meetbare inter-individuele variatie in gedragsmatige indicaties van cognitieve routinisering. Variatie is alomtegenwoordig, maar niet willekeurig. Als we tot accurate theorieën over de cognitieve representatie van taal willen komen, is het van belang dat we in kaart brengen welke factoren de variatie tussen taalgebruikers bepalen, en dit vereist dat we ons niet beperken tot geaggregeerde data, maar inzoomen op het niveau van individuen.

Afgezien van inter-individuele variatie, gaven mijn data ook blijk van intra-individuele variatie. Dit wijst eveneens op het dynamische karakter van mentale representaties van taal. Psycholinguïstische taken die inzicht trachten te krijgen in de mate waarin een taalelement verankerd is in iemands taalkennis drukken dit doorgaans uit in een enkele waarde (bijv. een reactietijd). Als cognitieve representaties de vorm aannemen van meer, of minder, compacte clusters bestaande uit *exemplars* die variëren in sterkte, dan levert een enkele waarde een incompleet beeld op. Om die reden pleit ik er niet alleen voor om aandacht te hebben voor variatie tussen mensen, maar ook om meerdere metingen per participant uit te voeren.

Tot besluit schets ik drie richtingen voor vervolgonderzoek die voortbouwen op het werk dat ik in dit proefschrift presenteer. Ten eerste stel ik voor om gepersonaliseerde maten verder te ontwikkelen. In mijn onderzoek heb ik de responses in de aanvultaak omgezet in een variabele die voor iedere participant afzonderlijk aangeeft of diegene het targetwoord wel of niet genoemd had (TargetMentioned). Dit bleek een waardevolle maat te zijn. Aangezien het een binaire variabele is, kan het echter geen recht doen aan graduele verschillen in de voorspelbaarheid van woorden. Ik doe suggesties voor manieren waarop het potentieel van gepersonaliseerde maten verkend kan worden.

Ten tweede stel ik voor om participanten gedurende enkele weken of maanden te volgen. Dit kan meer inzicht verschaffen in de effecten van gebruiksfrequentie op verwerkingssnelheid en vertrouwdheidsoordelen. Het is duidelijk dat frequentie van grote invloed is. Het is nog niet zo helder of het uitmaakt hoe recent een constructie is gebruikt; of het uitmaakt of je een constructie één of twee keer hebt gebruikt die dag; en of dit bij laagfrequente constructies anders uitpakt dan bij hoogfrequente.

Ten derde stel ik voor ik om, naast lexicaal specifieke constructies (zoals woorden en woordcombinaties), ook (gedeeltelijk) schematische constructies te onderzoeken. Volgens gebruiksgebaseerde theorieën komen alle mentale representaties van taal voort uit ervaringen met taal, waarbij gebruik wordt gemaakt van algemene cognitieve vaardigheden zoals patroonherkenning, *chunking,* categorisatie, en schematisering. Als (gedeeltelijk) schematische constructies gevormd worden door ervaringen met taal, dan zou ook hierbij inter- en intra-individuele variatie te verwachten zijn. Wel is het zo dat schematische constructies doorgaans een algemenere betekenis hebben, in een groter aantal contexten gebruikt worden, en een hogere gebruiksfrequentie hebben dan lexicaal specifieke constructies. Mogelijk doet zich hierdoor minder inter- en intra-individuele variatie voor. Waar ook rekening mee dient te worden gehouden is dat taalgebruikers onderling kunnen verschillen in cognitieve vermogens, zoals taalanalytisch vermogen, statistisch leervermogen, fluïde intelligentie, en cognitieve motivatie (Dąbrowska 2018; Misyak & Christiansen 2012). Zowel ervaringen met taal, als cognitieve vermogens lijken van invloed te zijn op het proces van schematiseren en kennis van grammaticale constructies. Er zijn aanwijzingen dat dit voor kennis van woordcombinaties niet op dezelfde manier geldt (Dąbrowska 2018). Mentale representaties van woorden, woordcombinaties, en abstractere patronen kunnen nog steeds opgevat worden als constructies die ontstaan uit ervaringen met taal in combinatie met algemene cognitieve vaardigheden, maar de mate waarin ze een beroep doen op bepaalde cognitieve en ervaringsgerichte factoren zou kunnen variëren. Ik hoop dat dit proefschrift een bijdrage levert aan deze onderzoeksagenda door te demonsteren dat het niet alleen mogelijk, maar ook zinvol is om rekening te houden met, en aandacht te schenken aan, inter- en intra-individuele variatie. Het zou mij zeer verheugen als mijn onderzoek taalkundigen weet te enthousiasmeren voor zo'n benadering.

Curriculum vitae

Véronique Verhagen (Eindhoven, 12 December 1985) obtained her gymnasium diploma from Lorentz Casimir Lyceum in 2003. She then completed the bachelor's program Linguistics and Intercultural Communication at Tilburg University, as well as the extracurricular, interdisciplinary honors program. Upon obtaining her bachelor's degree in 2006, she was awarded an Excellence Scholarship. Subsequently, she studied at Venice International University and took additional courses at Tilburg University. After that, she completed the research master's program in Language and Communication (Tilburg University and Radboud University Nijmegen), with a specialization in cognitive linguistics. For her thesis on individual differences in entrenchment of multi-words units, she received the Tilburg University research master's thesis award. After her graduation, she started as a PhD candidate at Tilburg University, supported by an NWO Promoties in de Geesteswetenschappen grant. In 2015 and 2016, she was appointed as a part-time lecturer at the department of Dutch Language and Culture at Leiden University. In 2017 and 2018, she taught a variety of courses in the department of Communication and Cognition at Tilburg University. As from January 2019, she is a lecturer in the Dutch teacher training program at Fontys University of Applied Sciences.