Véronique Verhagen*, Maria Mos, Joost Schilperoord and Ad Backus

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

https://doi.org/10.1515/ling-2018-0036

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 article, 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.

Keywords: individual variation, multiword units, metalinguistic judgments, linguistic representations, usage-based linguistics

***Corresponding author: Véronique Verhagen,** Department of Culture Studies, Tilburg University, D 418, Postbus 90153, 5000 LE Tilburg, The Netherlands,
E-mail: v.a.y.verhagen@tilburguniversity.edu
Maria Mos, Department of Communication and Information Sciences, Tilburg University, D 411, Postbus 90153, 5000 LE Tilburg, The Netherlands, E-mail: maria.mos@tilburguniversity.edu
Joost Schilperoord, Department of Communication and Information Sciences, Tilburg University, D 431, Postbus 90153, 5000 LE Tilburg, The Netherlands,
E-mail: j.schilperoord@tilburguniversity.edu
Ad Backus, Department of Culture Studies, Tilburg University, D 212, Postbus 90153, 5000 LE Tilburg, The Netherlands, E-mail: a.m.backus@tilburguniversity.edu

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 and Fedorenko 2010, 2013; Wasow and 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, Clark 1973 on the language-as-fixed-effect fallacy, and Baker and 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 and Fedorenko 2010, 2013; Wasow and 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 and 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" (Featherston 2007: 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" (Featherston 2007: 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 and 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 and 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.[1]

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 and 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 and 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 and 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 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

1 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).

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 and Schrauf 2009; 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 and 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/.[2] 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

2 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".

/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 and Gries 2009; Hashemi and 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 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 and Häussler 2010; Langsford et al. 2018; Preston and 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. 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 and Alexopoulou 2001; Meng and Bader 2000; Sorace 2000; Schütze 1996; Sprouse and Almeida 2012; Theakston 2004) to formulaic language (e.g. Ellis and Simpson-Vlach 2009), collocations and constructions (Granger 1998; Gries and 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.

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). 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 and Snider 2010; Verhagen et al. 2018; Tremblay and Tucker 2011), and we expect usage frequency to be reflected in familiarity ratings (cf. Balota et al. 2001; Popiel and 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 and Peter 1984; Jiang and 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; Ellis 1991; Johnson et al. 1996; Tabatabaei and 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 and 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 and 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 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 and 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 judgement, 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 and 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 and 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 and 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 Method

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 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 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

3.2 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 ($F(3, 87) = 0.20$, $p = 0.89$) or gender ($\chi^2(3) = 1.83$, $p = 0.63$). 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 ($\chi^2(6) = 3.57$, $p = 0.73$).

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 14,688) in a subset of the corpus SoNaR consisting of approximately 195.6 million words.[3] The phrases and the frequency data can be found in Appendices 1 and 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 (Verhagen and Mos 2016), 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.

3.4 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.[4] 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 and 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

3 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).

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

were marked by the words 'Disagree' and 'Agree' (*Oneens* and *Eens*). Participants were shown one example. After that, the experiment started.

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 ($F(3, 87) = 0.28$, $p = 0.84$). After four days, people are not expected to be able to recall the exact scores they assigned to each of the 79 stimuli.

3.5 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 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 and Häussler 2010; Schütze and 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 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 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).[5] To give an example, the phrase *naar huis* occurred 14,688

5 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 and Jaeger 2016). These metrics are not taken into account in the present study, as we do not

times, which corresponds to a log-transformed frequency score of 1.88. The lemma frequency of the noun, which encompasses occurrences of *huizen, huisje, huisjes* in addition to *huis*, amounts to 84,918 instances. This corresponds to a log-transformed frequency score of 2.64. Figure 1 shows the positions of the stimuli on the phrase frequency scale and the lemma frequency scale; Appendix 2 lists for all stimuli the raw and the log-transformed frequencies. As can be observed from Figure 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).

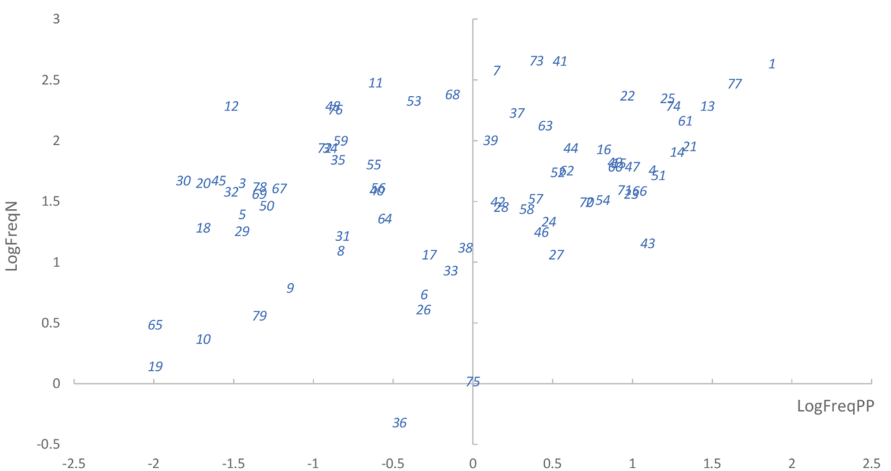

Figure 1: Scatterplot of the relationship between the log-transformed corpus frequency per million words of the PP and that of the N ($r = 0.39$). The numbers 1 to 79 identify the individual stimuli (see Appendices).

examine processing costs. We do so in another paper, in which we examine individual differences in experiences, expectations, and processing speed (Verhagen et al. 2018).

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 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 3.5, 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. 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).[6] If the latter is significantly higher than the former, the fourth hypothesis is confirmed.

6 Field (2013: 287) describes how one can test by means of a *t*-statistic (Chen and Popovich 2002) whether a difference between two dependent correlations from the same sample is

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.

4 Results[7]

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

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:

$$t_{\text{Difference}} = (r_{xy} - r_{zy}) * \sqrt{}(((n-3)(1+r_{xz}))/(2(1-r^2_{xy}-r^2_{xz}-r^2_{zy}+2*r_{xy}*r_{xz}*r_{zy})))$$

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 ($p < 0.05$) and 2.64 ($p < 0.01$).

7 The datasets and the scripts we used to analyze the data are available in DataverseNL at https://hdl.handle.net/10411/SUAUNY.

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. 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 2. The variance explained by this model is 57% ($R^2m = 0.36$, $R^2c = 0.57$).[8]

Table 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 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 2. This phrase frequency effect was found both in Likert and ME ratings, at Time 1 as well Time 2.

8 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).

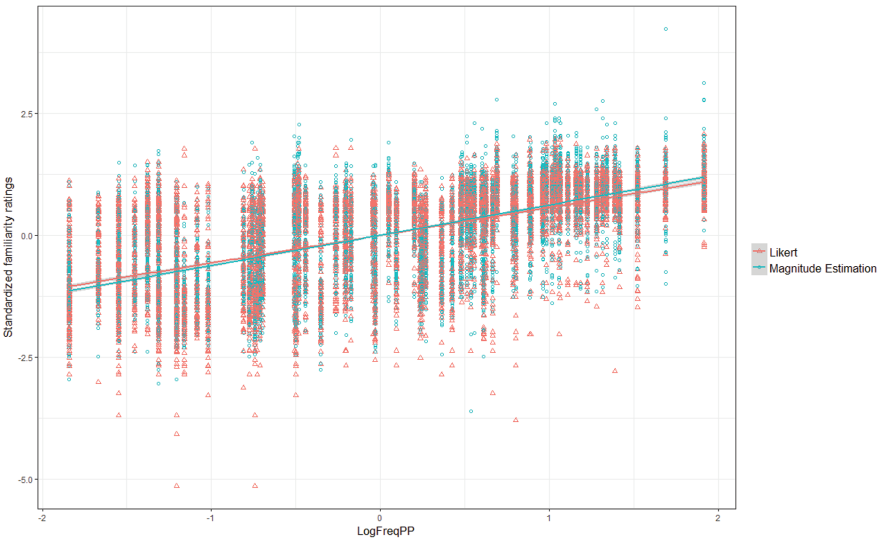

Figure 2: Scatterplot of the log-transformed corpus frequency per million words of the PP and the standardized familiarity ratings, split up according to whether the ratings were expressed on a 7-point Likert scale or a Magnitude Estimation scale. Each circle/triangle represents one observation; the lines represent linear regression lines with a 95% confidence interval around it.

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 2 quantify per item the amount of variation in judgment across participants. Figure 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.

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

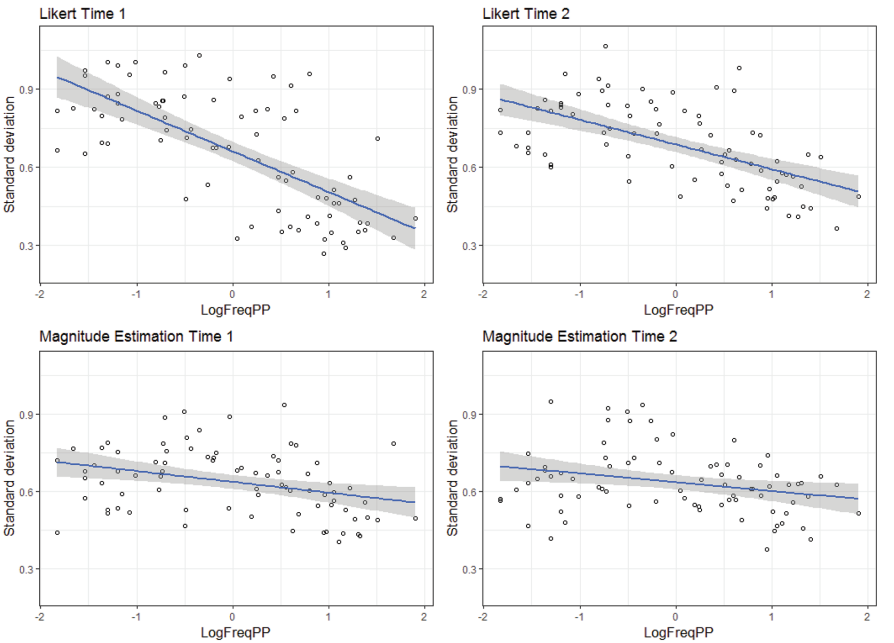

Figure 3: Scatterplots of the standard deviations in relation to the log-transformed corpus frequency per million words of the PP. The lines represent linear regression lines with a 95% confidence interval around it.

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).

Table 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	0.97	*0.96, 0.98*
Likert	ME	0.96	*0.94, 0.97*
ME	Likert	0.98	*0.97, 0.98*
ME	ME	0.98	*0.97, 0.99*

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 −0.31 to 0.90, with a mean correlation of 0.70 ($SD = 0.20$). The four conditions do not differ significantly in terms of intra-individual stability ($H(3) = 4.76$, $p = 0.19$). If anything, the ME-ME condition yields slightly more stable judgments than the other conditions, as can be observed from Table 4 and Figure 4.

Table 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	0.67 (0.27)	−0.31–0.87
Likert	ME	0.66 (0.21)	−0.01–0.86
ME	Likert	0.72 (0.14)	0.38–0.87
ME	ME	0.76 (0.11)	0.45–0.90

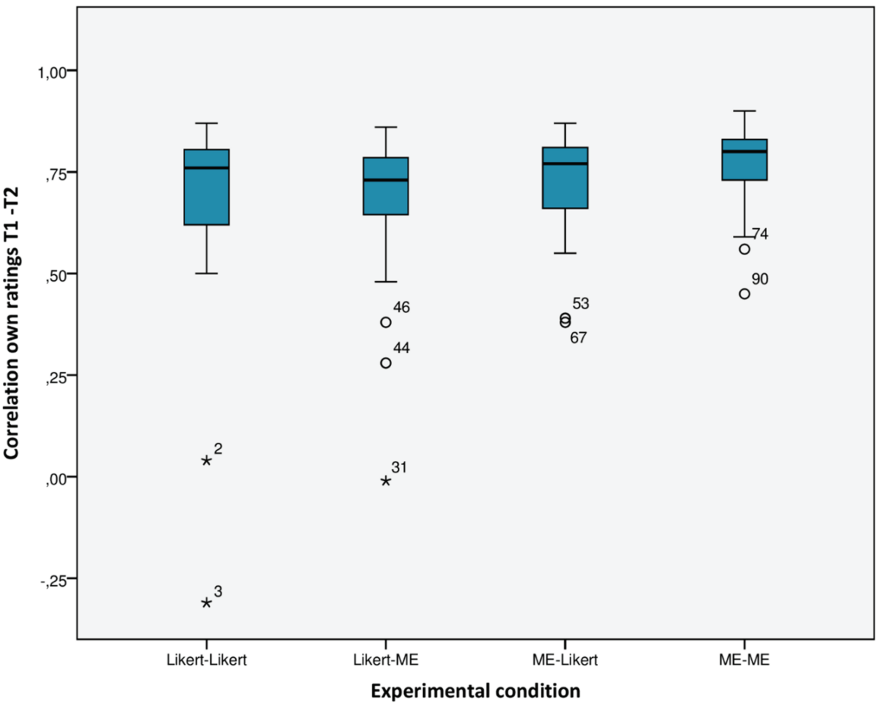

Figure 4: Boxplot of participants' correlation of their own standardized ratings (Pearson's *r*, Time 1 – Time 2).

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 ($r < 0.20$). Two of them were part of the Likert-Likert group; one of them belonged to the Likert-ME group.[9] 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 0.20 to 0.50; 34.1% (N = 31) had scores ranging from 0.51 to 0.75; 54.9% (N = 50) had scores ranging from 0.76 to 0.90. Still, none of the participants is as stable in their ratings as the aggregated ratings presented in Table 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.

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 3.5 and 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.[10] The resulting model is summarized in Table 5.

9 Low self-correlation scores are not related to educational background. The three participants with self-correlation scores below 0.20 had intermediate vocational education, higher vocational education, and higher education. As regards the group with self-correlation scores ranging from 0.20 to 0.49, one participant had intermediate vocational education, and the others had a tertiary education degree.

10 As was reported in Section 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

Table 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). 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 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 5).

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

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).

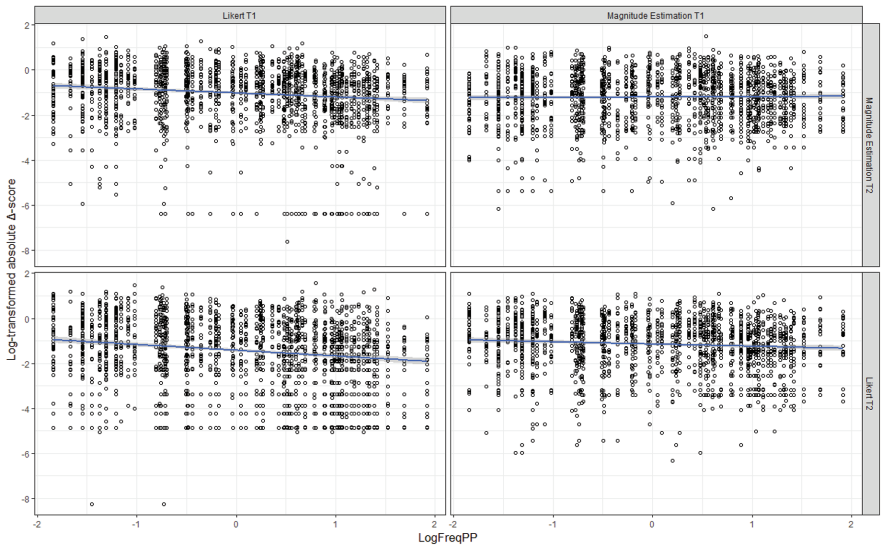

Figure 5: Scatterplot of the log-transformed corpus frequency per million words of the PP and the log-transformed absolute Δ-scores, per experimental condition. Each circle represents one observation; the lines represent linear regression lines with a 95% confidence interval around them.
Note: The lower 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.

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.[11]

11 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

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 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

---

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.

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.

Acknowledgements: We thank Carleen Baas for her help in collecting the data, and Martijn Goudbeek for his helpful comments and suggestions on this manuscript.

Funding: This work was supported by an NWO grant (The Netherlands Organization for Scientific Research, project number 322-89-004).

Appendix 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

(*continued*)

(*continued*)

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 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 N	Log FreqN	Freq PP	Log FreqPP	Likert M (SD)		ME M (SD)		Likert M (SD)		ME M (SD)	
1	naar huis	84,918	2.64	14,688	1.88	0.94	(0.41)	1.17	(0.50)	1.02	(0.50)	1.36	(0.51)
77	op school	58,222	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	37,893	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	17,713	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	28,615	2.17	4221	1.33	0.86	(0.36)	0.93	(0.56)	0.89	(0.66)	1.06	(0.57)
14	op vakantie	15,864	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	37,927	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	44,194	2.35	3259	1.22	0.59	(0.46)	0.94	(0.49)	0.77	(0.41)	0.74	(0.62)

(*continued*)

(*continued*)

						Time 1				Time 2			
		Freq N	Log FreqN	Freq PP	Log FreqPP	Likert M (SD)		ME M (SD)		Likert M (SD)		ME M (SD)	
51	in de tuin	10,213	1.72	2860	1.17	0.65	(0.58)	0.60	(0.60)	0.64	(0.57)	0.83	(0.55)
4	op papier	11,249	1.76	2606	1.12	0.80	(0.30)	0.82	(0.53)	0.75	(0.42)	0.75	(0.61)
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	12,003	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	46,260	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	12,870	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	12,053	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	13,022	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	16,681	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	17,185	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	11,205	1.76	756	0.59	0.19	(0.76)	0.14	(0.77)	0.05	(0.99)	0.18	(0.66)
41	onder elkaar	89,055	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	10,856	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	26,189	2.13	558	0.46	0.24	(0.75)	0.08	(0.94)	0.22	(0.54)	0.23	(0.69)

(*continued*)

(*continued*)

						Time 1				Time 2			
		Freq N	Log FreqN	Freq PP	Log FreqPP	Likert M (SD)		ME M (SD)		Likert M (SD)		ME M (SD)	
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	89,997	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	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	33,283	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	74,296	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	19,931	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)
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	47,205	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	42,001	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	59,244	2.48	47	−0.61	−1.36	(0.93)	−1.30	(0.47)	−1.30	(0.66)	−1.17	(0.69)

(*continued*)

(*continued*)

						Time 1				Time 2			
		Freq N	Log FreqN	Freq PP	Log FreqPP	Likert M (SD)		ME M (SD)		Likert M (SD)		ME M (SD)	
55	in de koffie	12,497	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)
59	uit de kunst	19,620	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	13,766	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	35,457	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	37,927	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	17,176	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	17,185	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)
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	37,927	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)

(*continued*)

(*continued*)

						Time 1				Time 2			
		Freq N	Log FreqN	Freq PP	Log FreqPP	Likert M (SD)		ME M (SD)		Likert M (SD)		ME M (SD)	
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. 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 ($\chi^2(1) = 86.64$, $p < 0.001$). After that, we added LogFreqN ($\chi^2(1) = 0.03$, $p = 0.87$) and the interaction term LogFreqPPxLogFreqN ($\chi^2(1) = 0.002$, $p = 0.96$), which did not improve model fit. We then proceeded with RatingScale ($\chi^2(1) = 0.0003$, $p = 0.99$), which did not improve model fit either. The interaction term RatingScaleT x LogFreqPP did contribute to the fit of the model ($\chi^2(2) = 21.79$, $p < 0.001$), as did RatingScale x LogFreqNP ($\chi^2(2) = 6.77$, $p < 0.05$). 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 ($\chi^2(2) = 0.00$, $p = 0.99$); Time x LogFreqPP ($\chi^2(1) = 0.01$, $p = 0.91$); Time x LogFreqN ($\chi^2(1) = 0.01$, $p = 0.91$)). Finally, PresentationOrder did not contribute to the goodness-of-fit ($\chi^2(1) = 1.27$, $p = 0.26$). Apart from the interaction term PresentationOrder x RatingScale ($\chi^2(2) = 7.05$, $p = 0.03$), none of the interactions of PresentationOrder and the other predictors in the model was found to improve model fit (PresentationOrder x LogFreqPP ($\chi^2(1) = 1.89$, $p = 0.17$); PresentationOrder x LogFreqN ($\chi^2(1) = 0.38$, $p = 0.54$); PresentationOrder x Time ($\chi^2(1) = 1.27$, $p = 0.26$); PresentationOrder x LogFreqPP x RatingScale ($\chi^2(2) = 5.41$, $p = 0.07$); PresentationOrder x LogFreqN x RatingScale ($\chi^2(2) = 0.46$, $p = 0.80$)). 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 ($\chi 2(3) = 90.21$, $p < 0.001$). 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 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 3.5). 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 ($\chi^2(1) = 32.92$, $p < 0.001$). After that, we added LogFreqN ($\chi^2(1) = 0.04$, $p = 0.84$). Given that LogFreqN did not improve model fit, we left out this predictor. We then proceeded with RatingScaleT1 ($\chi^2(1) = 0.15$, $p = 0.70$) and RatingScaleT2 ($\chi^2(1) = 2.39$, $p = 0.12$),

neither of which improved model fit. The interaction term RatingScaleT1 x RatingScaleT2 did not contribute to the fit of the model fit either ($\chi^2(3) = 6.67$, $p = 0.08$). The interaction term RatingScaleT1 x LogFreqPP did improve model fit ($\chi^2(2) = 40.94$, $p < 0.001$), as did RatingScaleT2 x LogFreqPP ($\chi^2(2) = 13.91$, $p < 0.001$). The three-way interaction RatingScaleT1 x RatingScaleT2 x LogFreqPP did not explain a significant portion of variance ($\chi^2(2) = 4.63$, $p = 0.10$). Finally, neither PresentationOrder ($\chi^2(1) = 0.27$, $p = 0.60$), nor any of the interactions of PresentationOrder and the other predictors in the model was found to improve model fit (PresentationOrder x LogFreqPP ($\chi^2(1) = 1.75$, $p = 0.19$); PresentationOrder x LogFreqPP x RatingScaleT1 ($\chi^2(2) = 2.52$, $p = 0.28$); PresentationOrder x LogFreqPP x RatingScaleT2 ($\chi^2(2) = 1.78$, $p = 0.41$)). 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 ($\chi2(2) = 12.96$, $p < 0.01$). 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 5.

References

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

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62. 67–82.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1). 1–27.

Ashton, Robert 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. Harald, Doug J. Davidson & Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412.

Baayen, R. Harald, Petar Milin, Dusica Filipovic Durdevic, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.

Baayen, R. Harald, Fabian Tomaschek, Susanne Gahl & Michael Ramscar. 2017. The Ecclesiastes principle in language change. In Marianne Hundt, Sandra Mollin & Simone E. Pfenninger (eds.), *The changing English language: Psycholinguistic perspectives*, 21–48. Cambridge: Cambridge University Press.

Backus, Ad. 2013. A usage-based approach to borrowability. In Eline Zenner & Gitte Kristiansen (eds.), *New perspectives on lexical borrowing*, 19–39. Berlin & Boston: De Gruyter Mouton.

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

Baker, Frank B. & Kim Seock-Ho. 2004. *Item response theory: Parameter estimation techniques*. New York: Dekker.

Balota, David A., Michael J. Cortese, Susan D. Sergent-Marshall, Daniel H. Spieler & Melvin J. Yap. 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General* 133(2). 283–316.

Balota, David A., Maura Pilotti & Michael J. Cortese. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* 29(4). 639–647.

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

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

Barlow, Michael & Suzanne Kemmer. 2000. *Usage-based models of language*. Cambridge: Cambridge University Press.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Bates, Douglas M., Martin Mächler, Benjamin M. Bolker & Steven C. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Birdsong, David. 1989. *Metalinguistic performance and interlinguistic competence*. New York: Springer.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82. 529–551.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Caldwell-Harris, Catherine, Jonathan Berant & Shimon Edelman. 2012. Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In Dagmar Divjak & Stephan Th. Gries (eds.), *Frequency effects in language representation*, 165–194. Berlin: Mouton de Gruyter.

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

Chen, Peter & Paula Popovich. 2002. *Correlation: Parametric and nonparametric measures*. Thousand Oaks, CA: Sage.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

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

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

Clark, Herbert 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.
Colman, Andrew M., Claire E. Norris & Carolyn C. Preston. 1997. Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports* 80(2). 355–362.
Crain, Stephen & Diane C. Lillo-Martin. 1999. *An Introduction to Linguistic Theory and Language Acquisition*. Malden: Blackwell.
Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16(3). 297–334.
Cumming, Geoff. 2014. The new statistics: Why and how. *Psychological Science* 25(1). 7–29.
Dąbrowska, Ewa. 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, Ewa. 2010. Naive v. expert competence: An empirical study of speaker intuitions. *The Linguistic Review* 27. 1–23.
De Bot, Kees & Robert Schrauf. 2009. *Language development over the lifespan*. New York: Routledge.
Eckert, Penelope. 1997. Age as a sociolinguistic variable. In Florian Coulmas (ed.), *Handbook of sociolinguistics*, 151–167. Oxford: Blackwell.
Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology* 41. 87–100.
Ellis, Nick 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, Nick. C. & Rita Simpson-Vlach. 2009. Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics and education. *Corpus Linguistics and Linguistic Theory* 5(1). 61–78.
Ellis, Rod. 1991. Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition* 13(2). 161–186.
Ellis, Rod. 2005. Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition* 27. 141–172.
Featherston, Sam. 2007. Data in generative grammar: The stick and the carrot. *Theoretical Linguistics* 33. 269–318.
Field, Andy. 2013. *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll*. Los Angeles, CA: Sage.
Field, Andy, Jeremy Miles & Zoë Field. 2012. *Discovering statistics using R*. London: Thousand Oaks.
Flynn, Suzanne. 1986. Production vs. comprehension: Differences in underlying competences. *Studies in Second Language Acquisition* 8. 135–164.
Foulkes, Paul. 2006. Phonological variation: A global perspective. In Bas Aarts & April McMahon (eds.), *The Handbook of English Linguistics*, 625–669. Oxford, UK: Blackwell.
Garrod, Simon & Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences* 8(1). 8–11.
Gernsbacher, Morton A. 1984. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* 113. 256–281.
Gibson, Edward & Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences* 14. 233–234.

Gibson, Edward & Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes* 28(1). 88–124.
Gilquin, Gaëtanelle & Stephan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26.
Goldinger, Stephen 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.
Goudbeek, Martijn, Daniel Swingley & Roel Smits. 2009. Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance* 35(6). 1913–1933.
Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis, and applications*, 145–160. Oxford: Oxford University Press.
Gries, Stefan Th. 2014. Quantitative corpus approaches to linguistic analysis: Seven or eight levels of resolution and the lessons they teach us. In Irma Taavitsainen, Merja Kytö, Claudia Claridge & Jeremy Smith (eds.), *Developments in English: Expanding electronic evidence*, 29–47. Cambridge: Cambridge University Press.
Gries, Stefan Th. 2015. Quantitative methods in linguistics. In James D. Wright (ed.), *International encyclopedia of the social and behavioral sciences*, 2nd edn, vol. 19, 725–732. Amsterdam: Elsevier.
Gries, Stefan Th. & Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7. 163–186.
Hashemi, Mohammad R. & Esmat Babaii. 2013. Mixed methods research: Toward new research designs in applied linguistics. *The Modern Language Journal* 97(4). 828–852.
Hintzman, Douglas L. 1986. "Schema abstraction" in a multiple-trace memory model. *Psychological Review* 93(4). 411–428.
Huettig, Falk. 2015. Four central questions about prediction in language processing. *Brain Research* 1626. 118–135.
Janda, Laura A. 2013. Quantitative methods in Cognitive Linguistics: An introduction. In Laura A. Janda (ed.), *Cognitive linguistics: The quantitative turn*, 1–32. Berlin: De Gruyter Mouton.
Jiang, X. Lu & Antonius Cillessen. 2005. Stability of continuous measures of sociometric status: A meta-analysis. *Developmental Review* 25(1). 1–25.
Johnson, Jacqueline S., Kenneth D. Shenkman, Elissa L. Newport & Douglas L. Medin. 1996. Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language* 35(3). 335–352.
Johnson, Paul C. D. 2014. Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution* 5. 944–946.
Juhasz, Barbara J., Yun-Hsuan Lai & Michelle L. Woodcock. 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.
Keller, Frank & Theodora Alexopoulou. 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.
Kertész, András, Monika Schwarz-Friesel & Manfred Consten. 2012. Introduction: Converging data sources in cognitive linguistics. *Language Sciences* 34(6). 651–655.

Kristiansen, Gitta & Rene Dirven. 2008. *Cognitive sociolinguistics: Language variation, cultural models, social systems*. Berlin: Mouton de Gruyter.
Kuhl, Patricia. 2000. A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America* 97(22). 11850–11857.
Labov, William. 1966. *The social stratification of English in New York City*. Washington: Center for Applied Linguistics.
Labov, William. 2001. *Principles of linguistic change. Vol. 2: Social factors*. Oxford: Blackwell.
Langsford, Steven, Amy Perfors, Andrew T. Hendrickson, Lauren A. Kennedy & Danielle J. Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: A Journal of General Linguistics* 3(1). 1–34.
Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. doi: 10.1016/j.cognition.2007.05.006
Linzen, Tal & Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40(6). 1382–1411.
Mandera, Pawel, Emmanuel Keuleers & Marc Brysbaert. 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, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.
Maxwell, Scott E., Ken Kelley & Joseph R. Rausch. 2008. Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology* 59. 537–563.
Meng, Michael & Markus Bader. 2000. Ungrammaticality detection and garden-path strength: Evidence for serial parsing. *Language and Cognitive Processes* 15. 615–666.
Mos, Maria, Antal van Den Bosch & Peter J. Berck. 2012. The predictive value of word-level perplexity in human sentence processing: A case study on fixed adjective-preposition constructions in Dutch. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency effects in language learning and processing*, 207–239. Berlijn: De Gruyter.
Nordquist, Dawn. 2009. Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory* 5(1). 105–130.
Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential speech and language technology for Dutch: Theory and applications of natural language processing*, 219–247. Dordrecht: Springer.
Paiva, Carlos, Eliane Barroso, Estela Carneseca, Cristiano de Pádua Souza, Felipe Thomé dos Santos, Rossana López & Bianca Sakamoto Ribeiro Paiva. 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.
Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam & Philadelphia: John Benjamins.
Popiel, Stephen J. & Ken McRae. 1988. The figurative and literal senses of idioms; or, all idioms are not used equally. *Journal of Psycholinguistic Research* 17. 475–487.
Preston, Carolyn C. & Andrew M. Colman. 2000. Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104(1). 1–15.

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

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

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

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

Schütze, Carson T. & Jon Sprouse. 2013. Judgment Data. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.

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

Seidenberg, Mark S. 1997. Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275. 1599–1603.

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

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

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

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

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

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

Street, James & Ewa Dąbrowska. 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, James & Ewa Dąbrowska. 2014. Lexically specific knowledge and individual differences in adult native speakers' processing of the English passive. *Applied Psycholinguistics* 35(1). 97–118.

Tabatabaei, Omid & Marzieh Dehghani. 2012. Assessing the reliability of grammaticality judgment tests. *Procedia – Social and Behavioral Sciences* 31. 173–182.

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

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of child language acquisition*. Cambridge: Harvard University Press.

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

VanGeest, Jonathan B., Matthew K. Wynia, Deborah S. Cummins & Ira B. Wilson. 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, Véronique & Maria Mos. 2016. Stability of familiarity judgments: Individual variation and the invariant bigger picture. *Cognitive Linguistics* 27(3). 307–344.

Verhagen, Véronique, Maria Mos, Ad Backus & Joost Schilperoord. 2018. Predictive language processing revealing usage-based variation. *Language and Cognition* 10(2). 329–373.
Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115. 1481–1496.
Wells, Justine B., Morten H. Christiansen, David S. Race, Daniel J. Acheson & Maryellen C. MacDonald. 2009. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology* 58. 250–271.
Weng, Li-jen. 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, Thomas & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87(2). 249–273.
Williams, Rihana & Robin Morris. 2004. Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology* 16. 312–339.
Wulff, Stefanie. 2009. Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory* 5(1). 131–159.