Language and Cognition **10** (2018), 329–373. doi:10.1017/langcog.2018.4

Predictive language processing revealing usage-based variation*

VÉRONIQUE VERHAGEN
Department of Culture Studies, Tilburg University

MARIA MOS
Department of Communication and Cognition, Tilburg University

AD BACKUS
Department of Culture Studies, Tilburg University

AND

JOOST SCHILPEROORD
Department of Communication and Cognition, Tilburg University

(*Received* 22 *May* 2017 – *Revised* 28 *February* 2018 – *Accepted* 12 *April* 2018)

ABSTRACT

While theories on predictive processing posit that predictions are based on one's prior experiences, experimental work has effectively ignored the fact that people differ from each other in their linguistic experiences and, consequently, in the predictions they generate. We examine usage-based variation by means of three groups of participants (recruiters, job-seekers, and people not (yet) looking for a job), two stimuli sets (word sequences characteristic of either job ads or news reports), and two experiments (a Completion task and a Voice Onset Time task). We show that differences in experiences with a particular register result in different expectations regarding word sequences characteristic of that register, thus pointing to differences in mental representations

[*] We thank Louis Onrust and Antal van den Bosch for their help in analyzing the corpus data, Sanneke Vermeulen for her help in collecting the experimental data, and Elaine Francis and two reviewers for their helpful comments and suggestions on this manuscript. Address for correspondence: 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

of language. Subsequently, we investigate to what extent different operationalizations of word predictability are accurate predictors of voice onset times. A measure of a participant's own expectations proves to be a significant predictor of processing speed over and above word predictability measures based on amalgamated data. These findings point to actual individual differences and highlight the merits of going beyond amalgamated data. We thus demonstrate that is it feasible to empirically assess the variation implied in usage-based theories, and we advocate exploiting this opportunity.

KEYWORDS: individual differences, surprisal, cloze probabilities, completion task, voice onset times.

1. Introduction

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

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

Crucially, in nearly all studies to date, the datasets providing word predictability measures come from different people than the datasets indicating performance in processing tasks, and that is a serious shortcoming.

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

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

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

In the following section, we discuss the concept of predictive processing in more detail. We describe how prediction in language processing is commonly investigated, focusing on the research design of those studies and the limitations. We then report on the outcomes of our study into variation in predictions and processing speed. The results show that there are meaningful differences to be detected between groups of speakers, and that a small

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

collection of data elicited from the participants themselves can be more informative than general corpus data. The prediction-based effects we observe are shown to be clearly influenced by differences in experience. On the basis of these findings, we argue that it is worthwhile to go beyond amalgamated data whenever prior experiences form a predictor in models of language processing and representation.

1.1. PREDICTION-BASED PROCESSING IN LANGUAGE

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

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

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

semantic, to phonological, orthographic, and perceptual – within their representation of a given context to predictively pre-activate information and facilitate the processing of new bottom-up inputs.

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

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

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

1.2. USAGE-BASED VARIATION IN PREDICTION-BASED PROCESSING

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

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

A comparison of groups of speakers to reveal usage-based variation appears to be a fruitful approach. Various studies indicate that people with different occupations (Dąbrowska, 2008; Gardner, Rothkopf, Lapan, & Lafferty, 1987; Street & Dąbrowska, 2010, 2014), from different social groups (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Caldwell-Harris, Berant, & Edelman, 2012), or with different amounts of training (Wells, Christiansen, Race, Acheson, & MacDonald, 2009) vary in the way they process particular words, phrases, or (partially) schematic constructions with which they can be expected to have different amounts of experience. To give an example, Caldwell-Harris and colleagues (2012) compared two groups with different prayer habits: Orthodox Jews and secular Jews. They administered a perceptual identification task in which phrases were briefly flashed on a

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

computer screen, one word immediately after the other. Participants were asked to report the words they saw, in the order in which they saw them. As expected, the two groups did not differ from each other in performance regarding the non-religious stimuli. On the religious phrases, by contrast, Orthodox Jews were found to be more accurate and to show stronger frequency effects than secular Jews. The participants who had greater experience with specific phrases could more easily match the brief, degraded input to a representation in long-term memory, recognize, and report it. Note, however, that these studies do not relate the performance on the experimental tasks to any other data from the participants themselves, and, with the exception of Street and Dąbrowska (2010, 2014), the researchers pay little attention to the degree of variation WITHIN each of the groups of participants.

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

2. Outline of the present research

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

To investigate this, we had three groups of participants – recruiters, job-seekers, and people not (yet) looking for a job – perform two tasks: a Completion task and a Voice Onset Time (VOT) task. In both tasks, we used two sets of stimuli: word sequences that typically occur in the domain of job hunting and word sequences that are characteristic of news reports. In the Completion task, the participants had to finish a given incomplete phrase (e.g., *goede contactuele* … 'good communication …'), listing all things that came to mind. In the VOT task, the participants were presented with the same cues, followed by a specific target word (e.g., *eigenschappen* 'skills'),

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

which they had to read aloud as quickly as possible. The voice onset times for this target word indicate how quickly it is processed in the given context.

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

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

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

2.1. PARTICIPANTS

122 native speakers of Dutch took part in this study. All of them had completed higher vocational or university education or were in the process of doing so. The participants belong to one of three groups. The first group,

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

labeled RECRUITERS, consists of 40 people (23 female, 17 male) who were working as a recruiter, intermediary, or HR adviser at the time of the experiment. Their ages range from 22 to 64, mean age 36.0 (SD = 10.0).

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

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

2.2. STIMULI

The stimuli consist of 35 word sequences characteristic of job advertisements and 35 word sequences characteristic of news reports. These word sequences were identified by using a Job ad corpus and the Twente News Corpus, and computing log-likelihood following the frequency profiling method of Rayson and Garside (2000). The Job ad corpus was composed by Textkernel, a company specializing in information extraction, web mining and semantic searching and matching in the Human Resources sector. All the job ads retrieved in the year 2011 (slightly over 1.36 million) were compiled, yielding a corpus of 488.41 million tokens. The Twente News Corpus (TwNC) is a corpus of comparable size (460.34 million tokens), comprising a number of national Dutch newspapers, teletext subtitling and autocues of broadcast news shows, and news data downloaded from the Internet (University of Twente, Human Media Interaction n.d.).[1] By means of the frequency profiling method we identified n-grams, ranging in length from three to ten words, whose occurrence frequency is statistically higher in one corpus than

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

another, thus appearing to be characteristic of the former (see Kilgarriff, 2001). In order to bypass an enormous amount of irrelevant sequences such as *Contract Soort Contract* and _ _ _ _ _ , which occur in the headers of the job ads, we applied the criterion that a sequence had to occur at least ten times in one corpus and two times in the other.

We selected sequences that met a number of additional requirements. A string had to end in a noun and it had to be comprehensible out of context. We only included *n*-grams that constitute a phrase, with clear syntactic boundaries. Sequences were also chosen in such a way that in the final set of stimuli all content words occur only once.[2] Furthermore, the selected sequences were to cover a range of values on two types of corpus-based measures: sequence frequency and surprisal of the final word in the sequence. With respect to the former, we took into account the frequency with which the sequence occurs in the specialized corpus (i.e., either the Job ad corpus or the News report corpus) as well as a corpus containing generic data, meant to reflect Dutch readers' overall experience, rather than one genre. We used a subset of the Dutch web corpus NLCOW14 (Schäfer & Bildhauer, 2012) as a generic corpus. The subset consisted of a random sample of 8 million sentences from NLCOW14, comprising in total 148 million words.

To obtain corpus-based surprisal estimates for the final word in the sequences, language models were trained on the generic corpus. These models were then used to determine the surprisal of the last word of the sequence (henceforth target word). Surprisal was estimated using a 7-gram modified Kneser–Ney algorithm as implemented in SRILM.[3]

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

2.3. PROCEDURE

The study consisted of a battery of tasks, administered in one session. Participants were tested individually in a quiet room. At the start of the session they were informed that the purpose of the study was to gain insight

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

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

into forms of communication in job ads and news reports and that they would be asked to read, complement, and judge short text fragments.

First, participants took part in the Completion task in which they had to complete the stimuli of which the final word had been omitted (see Section 3.1). After that, they filled out a questionnaire regarding demographic variables (age, gender, language background) and two short attention-demanding, arithmetic distractor tasks created using the Qualtrics software program. These tasks distracted participants from the word sequences that they had encountered in the Completion task and were about to see again in the Voice Onset Time experiment. After that, the VOT experiment started. In this task, participants were shown an incomplete stimulus (i.e., the last word was omitted), and then they saw the final word. They read aloud this target word as quickly as possible (see Section 4.1 for more details).

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

3. Experiment 1: completion task

3.1. METHOD

3.1.1. *Materials*

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

3.1.2. *Procedure*

Participants were informed that they were about to see a series of short text fragments. They were instructed to read them out loud and complete them by naming all appropriate complements that immediately come to mind. For this, they were given five seconds per trial. It was emphasized that there is no one correct answer. In order to reduce the risk of chaining (i.e., responding with associations based on a previous response rather than responding to the cue; see De Deyne & Storms, 2008; McEvoy & Nelson, 1982), participants were shown three examples in which the cue was repeated in every response (e.g., cue: *een kopje* … 'a cup of …', responses: *een kopje koffie, een kopje thee,*

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

een kopje suiker 'a cup of coffee, a cup of tea, a cup of sugar'). In this way, we prompted participants to repeat the cue every time, thus minimizing the risk of chaining.

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

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

3.1.3. *Scoring of responses*

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

By means of stereotypy points (see Fitzpatrick, Playfoot, Wray, & Wright, 2015) we quantified how similar each participant's responses are to the complements observed in the specialized corpora. The nominal complements that occurred in the corpus in question were assigned percentages that reflect

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 1. *Mean number of responses participants gave per cue; standard deviations between parentheses*

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

the relative frequency.[4] The sequence *40 uur per* '40 hours per', for example, was always followed by the word *week* 'week' in the Job ad corpus. Therefore, the response *week* was awarded 100 points; all other responses received zero points. In contrast, the sequence *kennis en* 'knowledge and' took seventy-three different nouns as continuations, a few of them occurring relatively often, and most occurring just a couple of times. Each response thus received a corresponding amount of points. For each stimulus, the points obtained by a participant were summed, yielding a stereotypy score ranging from 0 to 100.[5]

3.1.4. *Statistical analyses*

By means of a mixed-effects logistic regression model (Jaeger, 2008), we investigated whether there are significant differences across groups of participants and sets of stimuli in the proportion of responses that

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

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

correspond to a complement observed in the specialized corpora. Mixed-models obviate the necessity of prior averaging over participants and/or items, enabling the researcher to model random subject and item effects (Jaeger 2008). Appendix V describes our implementation of this statistical technique.

3.2. RESULTS

For each stimulus, participants obtained a stereotypy score that quantifies how similar their responses are to the complements observed in the specialized corpora. Table 2 presents the average scores of each of the groups on the two types of stimuli.

The average scores in Table 2 mask variation across participants within each of the groups (as indicated by the standard deviations) and variation across items within each of the two sets of stimuli. Figure 1 visualizes for each participant the mean stereotypy score on News report items and the mean stereotypy score on Job ad items. It thus sketches the extent to which scores on the two item types differ, as well as the extent to which participants within a group differ from each other. Figure 2 portrays these differences in another manner; it visualizes for each participant the difference in stereotypy scores on the two types of stimuli. The majority of the Recruiters obtained a higher stereotypy score on Job ad stimuli than on News report stimuli, as evidenced by the Recruiters' marks above the zero line. For the vast majority of the Inexperienced participants it is exactly the other way around: their marks are predominantly located below zero. The Job-seekers show a more varied pattern, with some participants scoring higher on Job ad items, some scoring higher on News report items, and some showing hardly any difference between their scores on the two sets of items.

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

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 2. *Mean stereotypy scores (on a 0–100 scale); standard deviations between parentheses*

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

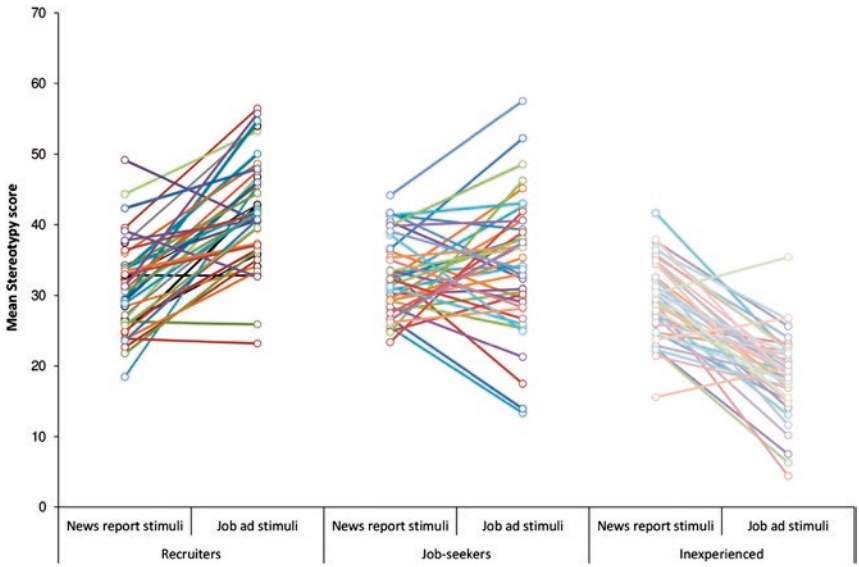

Fig. 1. Mean stereotypy score on the two types of stimuli for each individual participant.

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

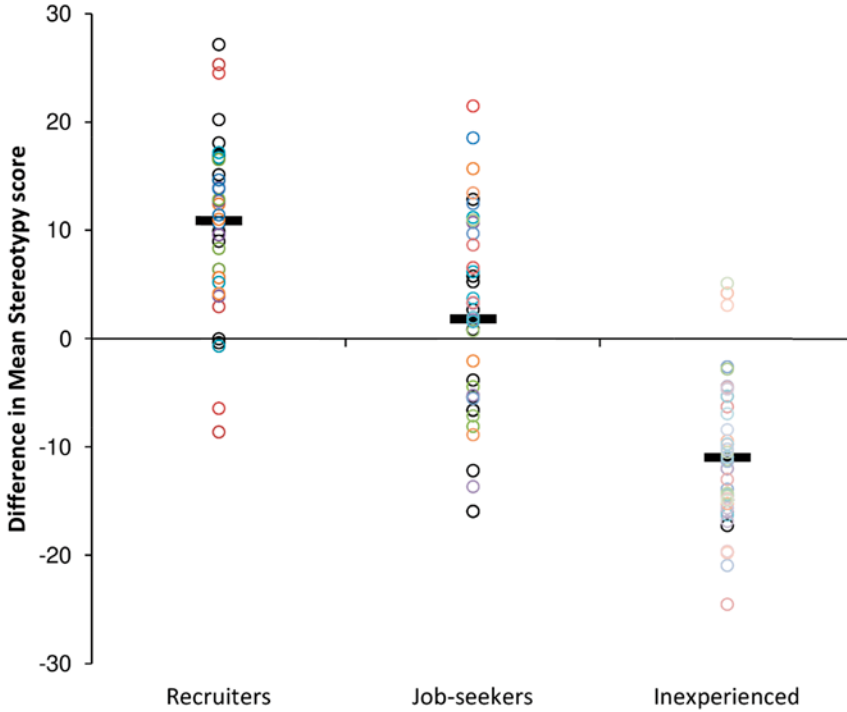

Fig. 2. The difference between the mean stereotypy score on Job ad stimuli and the mean stereotypy score on News report stimuli for each individual participant; black bars show each group's mean difference. A circle below zero indicates that that participant obtained higher stereotypy scores on News report stimuli than on Job ad stimuli.

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

3.3. DISCUSSION

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

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

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

4. Experiment 2: Voice Onset Time task

4.1. METHOD

4.1.1. *Materials*

The set of stimuli comprised the same 70 experimental items as the Completion task (35 Job ad word sequences and 35 News report word

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

sequences, described in Section 2.2) plus 17 filler items. The fillers were of the same type as the experimental items (i.e., (PREPOSITION) (ARTICLE) ADJECTIVE NOUN) and consisted of words unrelated to these items (e.g., *het prachtige uitzicht* 'the beautiful view'). The stimuli were randomized once. The presentation order was the same for all participants, to ensure that any differences between participants' responses are not caused by differences in stimulus order.

4.1.2. *Procedure*

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

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

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

All participants performed the task individually in a quiet room. The Inexperienced group was made up of students who were tested in sound-attenuated booths at the university. The Recruiters and Job-seekers were tested in rooms that were quiet, but not as free from distractions as the booths. This appears to have influenced reaction times: the Inexperienced participants responded considerably faster than the other groups (see Section 4.2). A by-subject random intercept in the mixed-effects models accounts for structural differences across participants in reaction times.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

4.1.3. *Data preparation and statistical analyses*

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

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

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

rLogFreq Word frequency has been shown to affect reading and naming times (Connine, Mullennix, Shernoff, & Yelen, 1990; Forster & Chambers, 1973; Kirsner, 1994; McDonald & Shillcock, 2003; Roland et al., 2012). It is a proxy for a word's familiarity and probability of occurrence without regard to context. We determined the frequency with which the target words (lemma search) occur in the generic corpus. This corpus comprised a wide range of texts, so as to reflect Dutch readers' overall experience, rather than one genre. The frequency counts were log-transformed. Word length and word frequency were correlated (r = –0.46), as was to be expected. Frequent words tend to have shorter linguistic forms (Zipf, 1935). We residualized

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

word frequency against word length, thus removing the collinearity from the model. The resulting predictor rLogFreq can be used to trace the influence of word frequency on VOTs once word length is taken into account.

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

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

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

GenericSurprisal The surprisal of the target word given the cue, estimated by language models trained on the generic corpus meant to reflect Dutch readers' overall experience (see Section 2.2 for more details).[6]

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

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

To give an idea of the number of times the target words were listed in the Completion task, Table 3 presents the mean percentage of target words mentioned by the participants in each of the groups.

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 3. *Mean percentage of targets words that had been mentioned by the participants in the Completion task; range between parentheses*

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

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

4.2. RESULTS

Table 4 presents for each group the mean voice onset time per item type. The Inexperienced participants were generally faster than the other groups, on both types of stimuli. This is likely due to factors irrelevant to our research questions: differences in experimental setting, in experience with participating in experiments, and in age. By-subject random intercepts account for such differences.[7] Of interest to us is the way the VOTs on the two types of items relate to each other, and the extent to which the VOTs can be predicted by measures of word predictability. These topics are discussed successively.

Table 4 shows that, on average, the Inexperienced participants responded faster to the News report stimuli than to the Job ad stimuli, while for the other groups it is just the other way around. Figures 3 and 4 visualize the pattern between the VOTs on the two types of items for each participant individually. For 80% of the Recruiters, the difference in mean VOTs on the two types of stimuli is negative, meaning that they were slightly faster to respond to Job ad stimuli than to News report stimuli. For 62.5% of the Job-seekers and 23.8% of the Inexperienced participants the difference score is below zero. Mixed-effects models fitted to the voice onset times (summarized

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 4. *Mean Voice Onset Times in seconds; standard deviations between parentheses*

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

in Table 4 and in Figures 3 and 4) revealed that the Inexperienced participants' data pattern is significantly different from the Recruiters' (β = –0.030, *SE* = 0.007, 99% CI: [–0.048, –0.011]) and the Job-seekers' (β = –0.019, *SE* = 0.005, 99% CI: [–0.034, –0.004]). That is, the fact that the Inexperienced participants tended to be faster on the News report items than on the Job ad items makes them differ significantly from both the Recruiters and the Job-seekers (see Appendix VI for more details).

What Figures 3 and 4 do not show is the degree of variation in VOTs across items within each of the two sets of stimuli. Every mark in Figure 3 averages over 35 items that differ from each other in word length, word frequency, and word predictability. By means of mixed-effects models, we examined to what extent these variables predict voice onset times, and whether there are effects of presentation order and block. We incrementally added predictors and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in voice onset times. A detailed description of this model selection procedure can be found in Appendix VI. The main outcomes are that the experimental design variable BLOCK and the interaction term PRESENTATIONORDER X BLOCK did not contribute to the fit of the model. The stimulus-related variables WORDLENGTH and RLOGFREQ did contribute. As for the word predictability measures, GENERICSURPRISAL did not improve model fit, but CLOZEPROBABILITY and TARGETMENTIONED did. While the interaction between RLOGFREQ and CLOZEPROBABILITY did not contribute to the fit of the model, the interaction between RLOGFREQ and TARGETMENTIONED did. None of the interactions of PRESENTATIONORDER and the other variables was found to improve goodness-of-fit. The resulting model is summarized in Table 5. The variance explained by this model is 60% (R^2m = .15, R^2c = .60).[8]

[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).

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

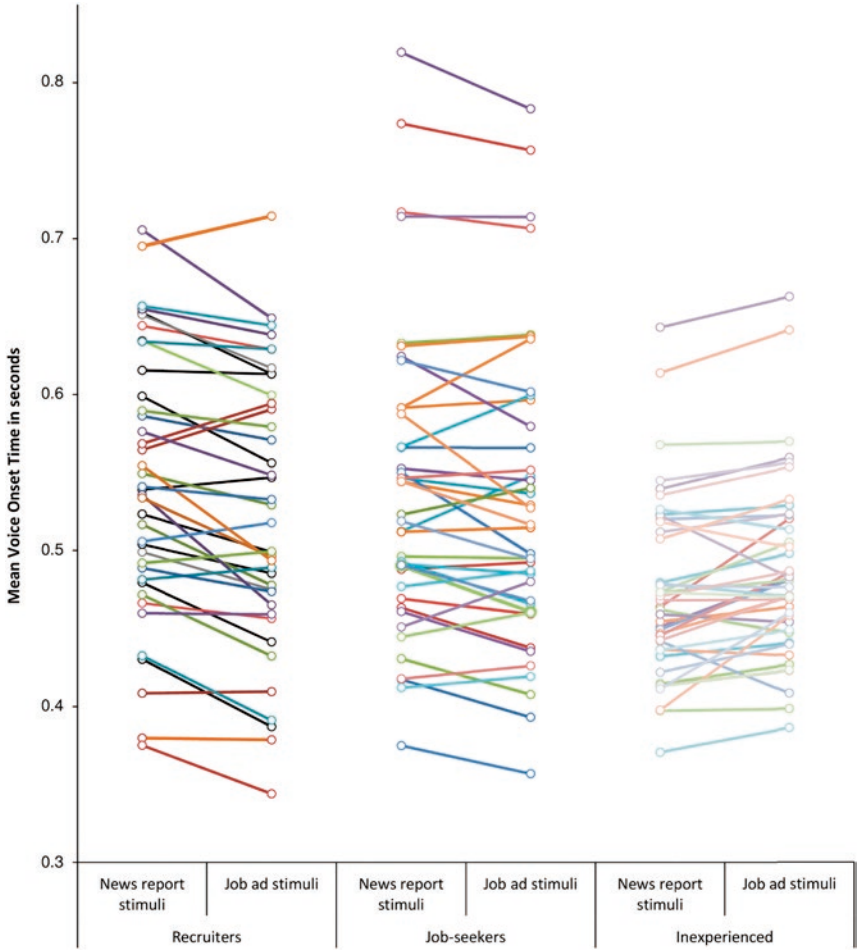

Fig. 3. Mean Voice Onset Time on the two types of stimuli for each individual participant.

Table 5 presents the outcomes when TARGET NOT MENTIONED is used as the reference condition. The intercept here represents the mean voice onset time when the target had not been mentioned by participants and all of the other predictors take their average value. A predictor's estimated coefficient indicates the change in voice onset times associated with every unit increase in that predictor. The estimated coefficient of rLOGFREQ, for instance, indicates that, when the target had not been mentioned and all other predictors take their average value, for every unit increase in residualized log frequency, voice onset times are 12 milliseconds faster.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

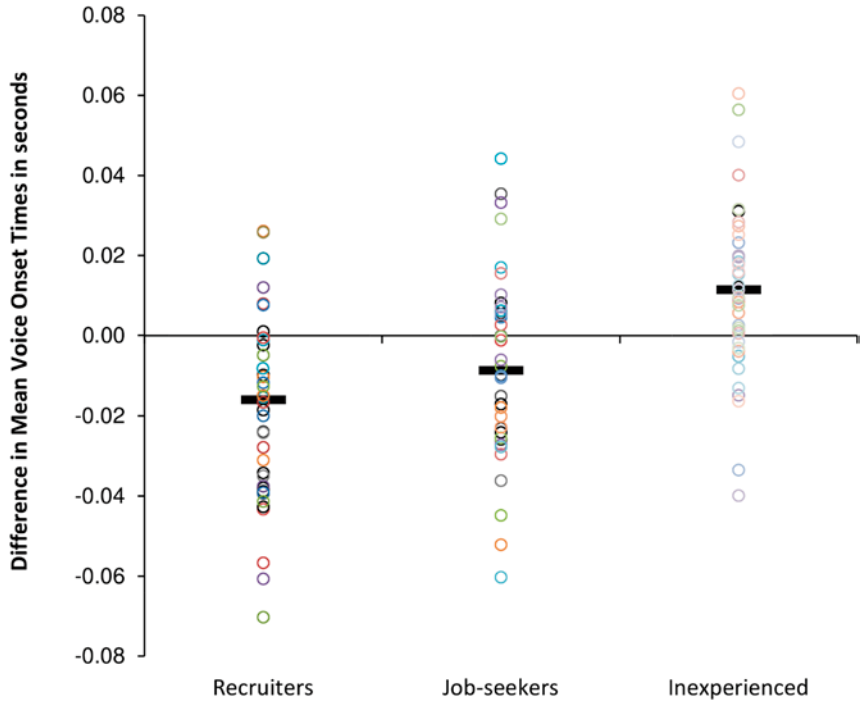

Fig. 4. The difference between the mean VOT on Job ad stimuli and the mean VOT on News report stimuli for each individual participant; black bars show each group's mean difference. A circle below zero indicates that that participant responded faster on Job ad stimuli than on News report stimuli.

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

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

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

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 5. *Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using 'Target not mentioned' as the reference condition*

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

NOTE: significance code: 0.01 '**'.

The effects of word frequency (rLogFreq) and TargetMentioned, and the interaction, are visualized in Figure 5. All along the frequency range, VOTs were significantly faster when the target had been mentioned by the participants in the preceding Completion task. The effect of TargetMentioned is more pronounced for lower-frequency items (the distance between the 'Target not mentioned' and the 'Target mentioned' line being larger on the left side than on the right side).

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

4.3. DISCUSSION

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

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

First of all, GenericSurprisal, which is the surprisal of the target word given the cue estimated by language models trained on the generic corpus,

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

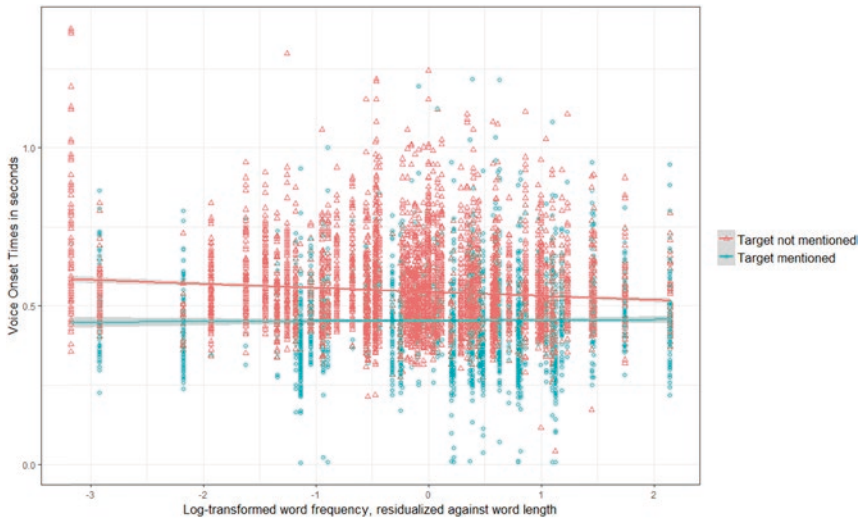

Fig. 5. Scatterplot of the log-transformed corpus frequency of the target word (lemma), residualized against word length, and the Voice Onset Times, split up according to whether or not the target word had been mentioned by a participant in the preceding Completion task. Each circle represents one observation; the lines represent linear regression lines with a 95% confidence interval around it.

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

The second finding is that ClozeProbability – a measure of word predictability based on the Completion task data of all 122 participants together – significantly predicted voice onset times. Target words with higher cloze probabilities were named faster. Combined, the first and the second finding indicate that general corpus data are too coarse an information

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

source for individual entrenchment, and that the total set of responses in a completion task from the participants themselves forms a better source of information.

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

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

Conversely, TargetMentioned explains variance that ClozeProbability does not account for. That is, participants were significantly faster to name the target if they had come up with this word to complete the phrase themselves approximately ten minutes earlier in the Completion task. This finding points to actual individual differences and

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

highlights the merits of going beyond amalgamated data. The fact that a measure of a participant's own predictions is a significant predictor of processing speed, over and above word predictability measures based on amalgamated data, had not yet been shown in lexical predictive processing research. It does fit in, more generally, with recent studies into the processing of schematic constructions in which individuals' scores from one experiment were found to correlate with their performance on another task (e.g., Misyak & Christiansen, 2012; Misyak, Christiansen, & Tomblin, 2010).

The fourth main finding is that the effect of TargetMentioned on voice onset times was stronger for lower-frequency than for higher-frequency items (the distance between the 'Target not mentioned' and the 'Target mentioned' line in Figure 5 being larger on the left side than on the right side). The high-frequency target words may be so familiar to the participants that they can process them quickly, regardless of whether or not they had pre-activated them. The processing of low-frequency items, on the other hand, clearly benefits from predictive pre-activation.

Fifth, corpus-based word frequency had no effect on VOTs when the target had been mentioned in the Completion task (i.e., $t = 0.13$ for rLogFreq; the 'Target mentioned' line in Figure 5 is virtually flat). In other words, predictive pre-activation facilitates processing to such an extent that word frequency no longer affects naming latency. When participants had not mentioned the target words in the Completion task, higher-frequency words elicited faster responses than lower-frequency words (in Table 5 rLogFreq is significant ($t = -2.58$); the 'Target not mentioned' line in Figure 5 descends).

5. General discussion

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

These findings point to a continuity with respect to observations in language acquisition research: the significance of individual differences and the merits

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

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

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

The data from the Completion task confirmed our hypotheses regarding the variation within and across groups in the predictions participants generate. On the News Report items, the groups did not differ significantly from each other in how likely participants were to name responses that correspond to the complements observed in the Twente News Corpus. On the Job ad stimuli, by contrast, all groups differed significantly from each other. The Recruiters' responses corresponded significantly more often to complements observed in the Job ad corpus than the Job-seekers' responses. The Job-seekers' responses, in turn, corresponded significantly more often to a complement in

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

the Job ad corpus than the responses of the Inexperienced participants. The responses thus reveal differences in participants' knowledge of multiword units which are related to their degree of experience with these word sequences.

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

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

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

Our study has focused on the processing of multiword units. Few linguists will deny there is individual variation in vocabulary inventories. In a usage-based approach to language learning and processing, there is no reason to assume that individual differences are restricted to concrete chunks such as words and phrases. One interesting next step, then, is to investigate to what extent similar differences can be observed for partially schematic or abstract patterns. Some of

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

these constructions (e.g., highly frequent patterns such as transitives) might be expected to show smaller differences, as exposure differs less substantially from person to person. However, recent studies point to individual differences in representations and processing of constructions that were commonly assumed to be shared by all adult native speakers of English (see Kemp, Mitchell, and Bryant, 2017, on the use of spelling rules for plural nouns and third-person singular present verbs in pseudo-words; Street and Dąbrowska, 2010, 2014, on passives and quantifiers). Our experimental set-up, which includes multiples tasks executed by the same participants, can also be used to investigate individual variation in processing abstract patterns and constructions.

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

REFERENCES

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

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

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

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

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

Barlow, M., & Kemmer, S. (2000). U*sage-based models of language*. Stanford, CA: CSLI Publications.

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

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

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

Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age – evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, **1**, 175–188.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: prediction takes precedence. *Cognition*, **136**, 135–149.
Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, **93**, 203–216.
Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.
Caldwell-Harris, C., Berant, J., & Edelman, Sh. (2012). Measuring mental entrenchment of phrases with perceptual identification, familiarity ratings, and corpus frequency statistics. In D. Divjak & S. Gries (Eds.), *Frequency effects in language representation* (pp. 165–194). Berlin: Mouton de Gruyter.
Carlsson, K., Petrovic, P., Skare, S., Petersoon, K. M., & Ingvar, M. (2000). Tickling expectations: neural processing in anticipation of a sensory stimulus. *Journal of Cognitive Neuroscience*, **12**(4), 691–703.
Chen, S., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**, 359–394.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, **36**(3), 181–204.
Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 1084–1096.
Croft, W. (2000). *Explaining language change: an evolutionary approach*. London: Longman.
Dąbrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: an empirical test of usage-based approaches to morphology. *Journal of Memory and Language*, **58**, 931–951.
Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, **16**(3), 437–474.
Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, **1084**(1), 89–103.
De Deyne, S., & Storms, G. (2008). Word associations: norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, **40**, 198–205.
DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, **8**(8), 1117–1121.
Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon: Association for Computational Linguistics.
Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from Chronolex. *Frontiers in Psychology*, **2**(306), 1–10.
Fitzpatrick, T., Playfoot, D., Wray, A., & Wright, M. (2015). Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics*, **36**, 23–50.
Forster, K., & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, **12**(6), 627–635.
Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, **5**, 475–494.
Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain & Language*, **140**, 1–11.
Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: finding a frequency-based component. *Memory and Cognition*, **15**, 24–28.
Goldberg, A. E. (2006). *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.
Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, **1626**, 118–135.
Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, **59**(4), 434–446.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, **5**, 944–946.
Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In R. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 135–168). Cambridge: Cambridge University Press.
Kemp, N., Mitchell, P., & Bryant, P. (2017). Simple morphological spelling rules are not always used: Individual differences in children and adults. *Applied Psycholinguistics*, **38**, 1071–1094.
Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**(1), 1–37.
Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283–312). San Diego, CA: Academic Press.
Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, **16**(1/2), 262–284.
Kristiansen, G., & Dirven, R. (2008). *Cognitive sociolinguistics: language variation, cultural models, social systems*. Berlin: Mouton de Gruyter.
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, **31**(1), 32–59.
Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: using our past to generate a future* (pp. 190–207). New York: Oxford University Press.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, **106**(3), 1126–1177.
Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: a usage-based analysis. *Cognitive Linguistics*, **20**(3), 481–507.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, **94**, 305–315.
McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, **14**(6), 648–652.
McEvoy, C. L., & Nelson, D. L. (1982). Category name and instance norms for 106 categories of various sizes. *American Journal of Psychology*, **95**, 581–634.
McNamara, T. P. (2005). *Semantic priming: perspectives from memory and word recognition*. Hove: Psychology Press.
Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: an individual differences study. *Language Learning*, **62**(1), 302–331.
Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential expectations: the role of prediction-based learning in language. *Topics in Cognitive Science*, **2**, 138–153.
Otten, M., & Van Berkum, J. (2008). Discourse-based anticipation during language processing: prediction or priming? *Discourse Processes*, **45**, 464–496.
Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: a critical review. *Psycholinguistic Bulletin*, **134**(3), 427–459.
R Core Team (2017). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: <http://www.R-project.org/>.
Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: implications for the E-Z reader model. *Journal of Experimental Psychology: Human Perception and Performance*, **30**(4), 720–732.
Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, 1–6.
Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore), 324–333.
Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, **122**, 267–279.
Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12).

Schmid, H.-J. (2015). A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association*, **3**, 1–27.
Simmons, W. K., Martin, A., & Barsalou, L. W. (2005) Pictures of appetizing foods activate gustatory cortices for taste and reward. *Cerebral Cortex*, **15**, 1602–1608.
Smith, N. J., & Levy, R. (2011). Cloze but no cigar: the complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1637–1642). Austin, TX: Cognitive Science Society.
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, **128**(3), 302–319.
Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. *Proceedings of the International Conference on Spoken Language Processing* (pp. 901–904). Denver, Colorado.
Street, J., & Dąbrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua*, **120**(8), 2080–2094.
Street, J., & Dabrowska, E. (2014). Lexically specific knowledge and individual differences in adult native speakers' processing of the English passive. *Applied Psycholinguistics*, **35**(1), 97–118.
Taylor, W. L. (1953). 'Cloze' procedure: a new tool for measuring readability. *Journalism Quarterly*, **30**, 415–433.
Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Traxler, M. J., & Foss, D. J. (2000). Effects of sentence constraint on priming in natural language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **26**(5), 1266–1282.
University of Twente, Human Media Interaction (n.d.). Twente News Corpus (TwNC): a multifaceted Dutch news corpus. Retrieved from <http://hmi.ewi.utwente.nl/TwNC/description>.
Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **31**, 443–467.
Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cognitive Psychology*, **58**, 250–271.
Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, **26**, 2506–2516.
Zipf, G. K. (1935). *The psychobiology of language: an introduction to dynamic philology*. Boston, MA: Houghton Mifflin Company.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

Appendix I

Job ad word sequences and corpus-based frequencies and surprisal estimates

The Job ad word sequences; base-10 logarithm of the frequency of occurrence per million words in the Job ad corpus and the NLCOW14-subset for the phrase as a whole and for the final word (lemma search); the surprisal of the final word based on data in NLCOW14-subset.

	Based on Job ad corpus	Based on NLCOW14-subset		
	LogFreq. phrase	LogFreq. phrase	Surprisal Final word	LogFreq. final word
1 40 uur per week	2.52	–0.40	41	0.92
2 voor meer informatie	2.36	0.37	84	1.33
3 kennis en ervaring	2.10	0.12	110	1.07
4 hoog in het vaandel	1.84	0.34	24	–0.32
5 werving en selectie	1.82	–0.54	119	0.63
6 een vast dienstverband	1.87	–0.77	332	–0.09
7 voor langere tijd	1.65	0.08	91	1.33
8 het eerste aanspreekpunt	1.48	–0.63	397	–0.15
9 goede contactuele eigenschappen	1.39	–1.22	339	0.82
10 bij gebleken geschiktheid	1.32	–1.06	217	–0.24
11 academisch werk- en denkniveau	1.00	–1.33	29	–0.85
12 een grote mate van zelfstandigheid	1.15	–0.99	46	0.07
13 in een hecht team	0.82	–1.57	119	1.08
14 een persoonlijk ontwikkelingsplan	0.55	–1.27	537	–0.71
15 een sterk analytisch vermogen	0.67	–1.69	208	0.89
16 met de mogelijkheid tot verlenging	0.50	–1.69	68	0.17
17 in de breedste zin van het woord	0.94	–0.04	9	0.96
18 met een afstand tot de arbeidsmarkt	0.05	–1.06	20	0.17
19 het geschetste profiel	0.24	–1.87	1546	0.58
20 in de meest uiteenlopende sectoren	0.39	–2.17	135	0.34
21 een vliegende start	0.10	–0.49	226	1.23
22 bewijs van goed gedrag	0.11	–0.99	71	1.17
23 conform de geldende CAO	–0.08	–1.87	151	0.20
24 met behoud van uitkering	–0.02	–0.51	45	0.47
25 bevoegd en bekwaam	–0.08	–1.39	247	0.04
26 een integrale benadering	–0.17	–0.81	342	0.83
27 naar aanleiding van de advertentie	–0.56	–2.17	96	0.28
28 eenvoudige administratieve werkzaamheden	–0.51	–1.87	919	0.82
29 een scherpe blik	–0.52	–1.17	447	0.86
30 buiten de geijkte paden	–0.90	–1.57	110	0.37
31 affiniteit met het onderwerp	–0.74	–1.69	112	0.84
32 een internationale speler van formaat	–1.24	–2.17	519	0.55
33 een flinke portie lef	–1.39	–2.17	344	0.07
34 met bewezen kwaliteiten	–1.17	–2.17	1586	0.59
35 een collegiale opstelling	–1.29	–2.17	13960	0.55

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

Appendix II

News report word sequences and corpus-based frequencies and surprisal estimates

The News report word sequences; base-10 logarithm of the frequency of occurrence per million words in the Twente News Corpus and the NLCOW14-subset for the phrase as a whole and for the final word (lemma search); the surprisal of the final word based on data in NLCOW14-subset.

	Based on News report corpus	Based on NLCOW14-subset		
	LogFreq. phrase	LogFreq. phrase	Surprisal final word	LogFreq. final word
36 de Tweede Kamer	1.94	0.38	144	0.31
37 wetenschap en techniek	1.87	–0.67	211	0.81
38 verkeer en vervoer	1.80	–0.52	169	0.57
39 in elk geval	1.71	0.84	52	0.65
40 in de Verenigde Staten	1.66	0.82	27	0.20
41 het openbaar ministerie	1.16	–0.32	264	0.22
42 de negentiende eeuw	1.05	–0.58	269	0.42
43 de raad van bestuur	1.04	–2.17	101	0.77
44 aan de andere kant	1.22	0.81	28	1.10
45 evenementen en manifestaties	1.47	–2.17	4662	0.00
46 het dagelijks leven	0.97	–0.20	213	1.50
47 op een gegeven moment	0.98	0.54	32	0.85
48 met terugwerkende kracht	0.58	0.26	77	1.03
49 in volle gang	0.66	0.05	96	0.60
50 een doorn in het oog	0.55	0.01	19	0.76
51 op geen enkele wijze	0.19	0.15	30	1.23
52 aan het begin van het seizoen	0.00	–0.38	15	0.74
53 de lokale bevolking	0.30	–0.25	329	0.78
54 het centrum van de stad	0.42	–0.59	50	1.02
55 correcties en aanvullingen	0.32	–1.17	189	0.06
56 de opvang van asielzoekers	–0.05	–1.09	149	0.32
57 de traditionele partijen	–0.42	–1.06	617	0.97
58 op last van de rechter	–0.35	–2.17	27	0.60
59 in de huidige situatie	–0.20	–0.22	56	0.99
60 een onafhankelijke commissie	–0.09	–0.83	382	0.65
61 een criminele afrekening	–0.83	–1.87	1486	–0.13
62 de koninklijke loge	–0.90	–1.87	1318	–0.54
63 een ingrijpende herstructurering	–0.90	–1.69	895	–0.01
64 op weg naar de top	–0.71	–1.22	32	0.93
65 in het belang van het kind	–0.63	–0.57	16	1.01
66 aan de vooravond van een revolutie	–1.36	–1.87	46	0.40
67 de uitkomsten van het rapport	–1.30	–1.69	73	0.78
68 met hernieuwde energie	–1.46	–1.17	262	1.03
69 een ongekende vrijheid	–1.38	–2.17	886	0.79
70 een luxe jacht	–1.46	–2.17	1092	0.35

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

Appendix III

Average Stereotypy Scores for the Job ad stimuli

Cue	Stereotypy Scores		
	Recruiters	Job-seekers	Inexperienced
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
1 40 uur per	97.5 (15.8)	97.5 (15.8)	90.5 (29.7)
2 voor meer	58.2 (48.1)	58.3 (48.1)	55.4 (48.6)
3 kennis en	21.0 (29.2)	12.9 (23.7)	6.3 (19.1)
4 hoog in het	90.0 (30.4)	82.5 (38.5)	66.7 (47.7)
5 werving en	96.7 (0.0)	84.6 (32.4)	27.6 (44.2)
6 een vast	15.8 (19.8)	13.7 (22.0)	5.2 (8.3)
7 voor langere	47.9 (43.7)	64.9 (38.0)	63.3 (40.4)
8 het eerste	2.4 (13.0)	0.2 (0.6)	0.0 (0.1)
9 goede contactuele	57.5 (50.1)	52.5 (50.6)	2.4 (15.4)
10 bij gebleken	74.8 (43.7)	29.9 (46.3)	2.4 (15.4)
11 academisch werk- en	85.0 (36.2)	57.5 (50.1)	0.0 (0.0)
12 een grote mate van	25.4 (32.2)	11.8 (25.5)	3.2 (14.3)
13 in een hecht	52.5 (50.6)	40.0 (49.6)	11.9 (32.8)
14 een persoonlijk	17.3 (23.4)	13.7 (21.9)	13.1 (21.5)
15 een sterk analytisch	95.0 (22.1)	80.0 (40.5)	66.7 (47.7)
16 met de mogelijkheid tot	33.9 (46.0)	22.0 (40.3)	1.0 (1.9)
17 in de breedste zin van het	100.0 (0.0)	95.0 (22.1)	78.6 (41.5)
18 met een afstand tot de	55.0 (50.4)	2.5 (15.8)	0.0 (0.0)
19 het geschetste	35.0 (48.3)	17.5 (38.5)	0.0 (0.0)
20 in de meest uiteenlopende	0.2 (0.4)	0.1 (0.3)	0.2 (0.4)
21 een vliegende	77.8 (38.3)	70.5 (43.2)	13.9 (34.2)
22 bewijs van goed	97.5 (15.8)	100.0 (0.0)	76.2 (43.1)
23 conform de geldende	13.9 (31.8)	5.3 (15.6)	4.1 (2.5)
24 met behoud van	25.8 (32.0)	9.7 (23.3)	0.0 (0.0)
25 bevoegd en	22.5 (42.3)	17.5 (38.5)	7.1 (26.1)
26 een integrale	12.9 (20.5)	13.4 (21.1)	0.2 (1.1)
27 naar aanleiding van de	9.4 (22.6)	6.5 (19.0)	0.0 (0.0)
28 eenvoudige administratieve	50.0 (50.6)	50.0 (50.6)	28.6 (45.7)
29 een scherpe	9.0 (9.7)	9.4 (8.5)	5.1 (8.2)
30 buiten de geijkte	56.1 (36.4)	48.6 (40.8)	4.3 (17.8)
31 affiniteit met het	13.4 (17.7)	10.3 (17.5)	8.8 (15.4)
32 een internationale speler van	2.5 (15.8)	7.5 (26.7)	0.0 (0.0)
33 een flinke portie	0.0 (0.0)	0.9 (5.6)	0.8 (5.4)
34 met bewezen	2.5 (15.8)	2.5 (15.8)	2.4(15.4)
35 een collegiale	14.1 (33.6)	11.8 (31.2)	0.0 (0.0)

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

Appendix IV

Average Stereotypy Scores for the News report stimuli

Cue	Stereotypy Scores		
	Recruiters	Job-seekers	Inexperienced
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
36 de Tweede Kamer	34.9 (18.2)	39.4 (16.9)	32.6 (17.2)
37 wetenschap en	5.3 (20.5)	8.0 (23.0)	10.6 (28.3)
38 verkeer en	7.5 (21.0)	15.9 (25.5)	2.7 (7.5)
39 in elk	73.1 (42.7)	87.7 (29.6)	65.0 (46.4)
40 in de Verenigde	86.5 (32.9)	96.3 (15.5)	94.1 (21.3)
41 het openbaar	21.9 (36.2)	22.3 (36.6)	14.7 (31.2)
42 de negentiende	72.8 (42.6)	63.1 (46.9)	50.8 (49.0)
43 de raad van	30.7 (13.8)	27.5 (18.5)	24.0 (18.9)
44 aan de andere	98.0 (0.6)	95.7 (15.5)	98.3 (0.9)
45 evenementen en	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
46 het dagelijks	34.4 (29.9)	46.1 (25.9)	38.5 (31.8)
47 op een gegeven	100.0 (0.0)	95.0 (22.1)	100.0 (0.0)
48 met terugwerkende	97.5 (15.8)	97.5 (15.8)	92.9 (26.1)
49 in volle	11.4 (21.1)	10.0 (18.5)	7.9 (14.2)
50 een doorn in het	95.0 (22.1)	97.5 (15.8)	90.5 (29.7)
51 op geen enkele	53.7 (31.4)	61.9 (29.2)	60.7 (32.7)
52 aan het begin van het	3.6 (12.5)	5.7 (13.5)	7.5 (19.2)
53 de lokale	7.0 (12.5)	5.5 (10.3)	7.8 (12.0)
54 het centrum van de	54.7 (47.6)	45.2 (48.1)	68.0 (43.5)
55 correcties en	22.5 (42.3)	25.0 (43.9)	7.1 (26.1)
56 de opvang van	8.3 (24.3)	4.3 (17.7)	6.6 (20.8)
57 de traditionele	3.6 (7.6)	4.1 (6.9)	1.2 (3.8)
58 op last van de	0.0 (0.0)	7.5 (26.7)	9.5 (29.7)
59 in de huidige	19.2 (17.1)	12.8 (16.1)	16.5 (16.4)
60 een onafhankelijke	5.2 (10.9)	3.4 (8.6)	7.0 (12.1)
61 een criminele	28.8 (44.5)	26.4 (43.3)	13.7 (33.9)
62 de koninklijke	13.6 (13.2)	11.6 (12.9)	17.8 (13.9)
63 een ingrijpende	5.5 (4.2)	6.3 (5.7)	4.8 (4.2)
64 op weg naar de	3.9 (13.9)	11.7 (24.7)	1.3 (8.5)
65 in het belang van het	3.0 (12.0)	1.9 (10.9)	1.6 (10.6)
66 aan de vooravond van een	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
67 de uitkomsten van het	63.6 (44.3)	77.4 (36.1)	62.5 (44.7)
68 met hernieuwde	12.5 (33.5)	10.0 (30.4)	2.4 (15.4)
69 een ongekende	2.2 (9.7)	1.4 (8.9)	0.0 (0.0)
70 een luxe	8.3 (8.8)	14.4 (13.8)	12.5 (18.1)

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

Appendix V

Mixed-effects logistic regression model fitted to the Completion task data

The stereotypy scores were not normally distributed. Therefore, it was not justified to fit a linear mixed-effects model. We used a mixed-effects logistic regression model (Jaeger, 2008) instead. Per response, we indicated whether or not it corresponded to a complement observed in the specialized corpora. By means of a mixed logit-model, we investigated whether there are significant differences across groups of participants and/or sets of stimuli in the proportion of responses that correspond to a complement in the specialized corpora. We fitted this model using the LMER function from the lme4 package in R (version 3.3.3; CRAN project; R Core Team, 2017). GROUP, ITEMTYPE, and their interaction were included as fixed effects, and participants and items as random effects. The fixed effects were standardized. Random intercepts and random slopes for participants and items were included to account for between-subject and between-item variation.[9]

A model with a full random effect structure was constructed following Barr, Levy, Scheepers, and Tily (2013). A comparison with the intercept-only model proved that the inclusion of the by-item random slope for GROUP and the by-participant random slope for ITEMTYPE was justified by the data ($\chi^2(7) = 174.83$, $p < .001$). Confidence intervals were estimated via parametric bootstrapping over 1,000 iterations (Bates, Mächler, Bolker, & Walker, 2015).

In order to obtain all relevant comparisons of the three groups and the two types of stimuli, we ran the model with different coding schemes and we report 99% confidence intervals (as opposed to the more common 95%) to correct for multiple comparisons. Since the groups were not expected to differ systematically in experience with News report word sequences, none of the groups forms a natural baseline in this respect. As for the Job ad stimuli, from a usage-based perspective, differences between Recruiters and Job-seekers are as interesting as differences between Job-seekers and Inexperienced participants, or Recruiters and Inexperienced participants. Therefore, we treatment-coded the factors, first using 'Recruiters' as the reference group for GROUP and 'Job ad stimuli' as the reference group for ITEMTYPE. The resulting model is summarized in Table 6. The intercept represents the proportion of the Recruiters' responses to the Job ad stimuli that correspond to a complement in the Job ad corpus. This proportion

[9] By-participant random slopes for GROUP were not included, as this was a between-participants factor; by-item random slopes for ITEMTYPE were not included, as this was a between-items factor.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 6. *Mixed-effects logistic regression model (family: binomial) fitted to the responses to the Completion task (0 = does not correspond to a complement in the specialized corpus; 1 = corresponds to a complement in the specialized corpus), using 'Recruiters–Job ad stimuli' as the reference condition*

	Estimate	Std. Error	*z*	99 % CI	
(Intercept)	0.56	0.43	1.31	*–0.54, 1.65*	
Itemtype_NewsReport	–0.56	0.60	–0.93	*–2.06, 0.97*	
Group_Jobseekers	–0.69	0.17	–4.09	*–1.11, –0.26*	**
Group_Inexperienced	–2.38	0.29	–8.30	*–3.11, –1.64*	**
Itemtype_NewsReport x Group_Jobseekers	0.91	0.21	4.36	*0.36, 1.46*	**
Itemtype_NewsReport x Group_Inexperienced	2.14	0.38	5.62	*1.15, 3.09*	**

NOTE: Significance code: 0.01 '**'.

does not differ significantly from the proportion of their responses to the News report items that correspond to a complement in the Twente News Corpus.

There are significant differences between the groups of participants on the Job ad stimuli. Both the Inexperienced participants and the Job-seekers have significantly lower proportions of responses to the Job ad stimuli that match a complement in the Job ad corpus than the Recruiters. The model also reveals that the difference between the proportions on the two types of stimuli is significantly different across groups.

To examine the remaining differences, we then used 'Job-seekers–Job ad stimuli' as the reference condition. The outcomes are summarized in Table 7. The proportion of the Job-seekers' responses to the Job ad items that correspond to a complement in the Job ad corpus does not differ significantly from the proportion of their responses to the News report items that match a complement in the Twente News Corpus. Furthermore, the outcomes show that the Job-seekers' responses to the Job ad stimuli were significantly more likely to correspond to a complement in the Job ad corpus than the responses of the Inexperienced participants. In addition, the model reveals that the difference between the proportions on the two types of stimuli is significantly different for the Inexperienced participants compared to the Job-seekers.

Finally, we used 'Inexperienced-News report stimuli' as the reference condition. The outcomes, summarized in Table 8, show that the proportion of the Inexperienced participants' responses to the Job ad items that correspond to a complement in the specialized corpus is not significantly different from the proportion of their responses to the News report items that match a complement in the specialized corpus. They also reveal that

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 7. *Mixed-effects logistic regression model (family: binomial) fitted to the responses to the Completion task (0 = does not correspond to a complement in the specialized corpus; 1 = corresponds to a complement in the specialized corpus), using 'Job-seekers–Job ad stimuli' as the reference condition*

	Estimate	Std. Error	*z*	99 % CI	
(Intercept)	–0.13	0.40	–0.32	*–1.13, 0.86*	
Itemtype_NewsReport	0.35	0.56	0.63	*–1.04, 1.72*	
Group_Inexperienced	–1.69	0.25	–6.78	*–2.34, –1.04*	**
Group_Recruiters	0.69	0.17	4.09	*0.25, 1.14*	**
Itemtype_NewsReport x Group_Inexperienced	1.23	0.32	3.79	*0.38, 2.07*	**
Itemtype_NewsReport x Group_Recruiters	–0.91	0.21	–4.36	*–1.43, –0.39*	**

NOTE: Significance code: 0.01 '**'.

TABLE 8. *Mixed-effects logistic regression model (family: binomial) fitted to the responses to the Completion task (0 = does not correspond to a complement in the specialized corpus; 1 = corresponds to a complement in the specialized corpus), using 'Inexperienced–News report stimuli' as the reference condition*

	Estimate	Std. Error	*z*	99 % CI	
(Intercept)	–0.24	0.45	–0.62	*–1.34, 0.84*	
Itemtype_JobAd	–1.58	0.64	–2.47	*–3.12, 0.01*	
Group_Jobseekers	0.46	0.23	1.98	*–0.11, 1.04*	
Group_Recruiters	0.24	0.27	0.88	*–0.44, 0.92*	
Itemtype_JobAd x Group_Jobseekers	1.23	0.32	3.79	*0.38, 2.04*	**
Itemtype_JobAd x Group_Recruiters	2.14	0.38	5.61	*1.14, 3.11*	**

NOTE: Significance code: 0.01 '**'.

the three groups do not differ significantly from each other in the proportion of responses to the News report stimuli that match a complement in the specialized corpus.

Appendix VI

Linear mixed-effects models fitted to the voice onset times (VOT task)

We fitted linear mixed-effects models (Baayen et al., 2008), using the LMER function from the lme4 package in R (version 3.3.2; CRAN project; R Core Team, 2017), to the Voice Onset Times. First, we investigated whether there are significant differences in VOTs across groups of participants and/or sets of stimuli, similar to our analysis of the stereotypy scores. Subsequently, we examined to what extent the VOTs can be predicted by word length, corpus-based word frequency, presentation order, and different measures of word predictability.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

In the first analysis, GROUP, ITEMTYPE, and their interaction were included as fixed effects, and participants and items as random effects. The fixed effects were standardized. We included random intercepts and slopes for participants and items to account for between-subject and between-item variation.[10]

A model with a full random effect structure was constructed following Barr et al. (2013). A comparison with the intercept-only model proved that the inclusion of the by-item random slope for GROUP and the by-participant random slope for ITEMTYPE was justified by the data ($\chi^2(7) = 34.34$, $p < .001$). The variance explained by this model is 59% ($R^2m = .04$, $R^2c = .59$).[11] Confidence intervals were estimated via parametric bootstrapping over 1,000 iterations (Bates et al., 2015).

In order to obtain all relevant comparisons of the three groups and the two types of stimuli, we ran the model with different coding schemes and we report 99% confidence intervals to correct for multiple comparisons. We treatment-coded the factors, first using 'Recruiters' as the reference group for GROUP and 'Job ad stimuli' as the reference group for ITEMTYPE. The resulting model is summarized in Table 9. The intercept represents the mean VOT of the Recruiters on the Job ad stimuli. Subsequently, we used 'Job-seekers–Job ad stimuli' as the reference condition (Table 10), and finally 'Inexperienced–News report stimuli' (Table 11).

The models reveal that none of the groups shows a significant difference between VOTs on the News report items and VOTs on the Job ad items. The Inexperienced do differ significantly from the Recruiters and Job-seekers in the relationship between the two sets of items. The majority of the Recruiters and the Job-seekers responded faster to the Job ad items than to the News report items (as evidenced by the Recruiters' and Job-seekers' marks below the zero line in Figure 4). For the vast majority of the Inexperienced participants it is just the other way around: they were faster on the News report stimuli compared to the Job add stimuli. The mixed-effects models indicate that the Inexperienced participants' data pattern is significantly different from the Recruiters' and the Job-seekers'.

In the second analysis, we investigated to what extent the VOTs can be predicted by various characteristics of the target words. We included the length

[10] By-participant random slopes for GROUP were not included, as this was a between-participants factor; by-item random slopes for ITEMTYPE were not included, as this was a between-items factor.

[11] 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).

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

TABLE 9. *Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using 'Recruiters–Job ad stimuli' as the reference condition*

	Estimate	Std. Error	*t*	99 % CI	
(Intercept)	0.522	0.017	30.27	*0.477, 0.566*	
Itemtype_NewsReport	0.020	0.016	1.24	*–0.024, 0.064*	
Group_Jobseekers	0.009	0.019	0.50	*–0.036, 0.057*	
Group_Inexperienced	–0.036	0.019	–1.93	*–0.085, 0.013*	
Itemtype_NewsReport x Group_Jobseekers	–0.011	0.006	–1.88	*–0.026, 0.004*	
Itemtype_NewsReport x Group_Inexperienced	–0.030	0.007	–4.12	*–0.048, –0.011*	**

NOTE: Significance code: 0.01 '**'.

TABLE 10. *Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using 'Job-seekers–Job ad stimuli' as the reference condition*

	Estimate	Std. Error	*t*	99 % CI	
(Intercept)	0.531	0.017	32.09	*0.488, 0.574*	
Itemtype_NewsReport	0.009	0.015	0.62	*–0.028, 0.047*	
Group_Recruiters	–0.009	0.019	–0.50	*–0.058, 0.040*	
Group_Inexperienced	–0.045	0.018	–2.47	*–0.094, 0.003*	
Itemtype_NewsReport x Group_Recruiters	0.011	0.006	1.88	*–0.004, 0.026*	
Itemtype_NewsReport x Group_Inexperienced	–0.019	0.005	–3.43	*–0.034, –0.004*	**

NOTE: Significance code: 0.01 '**'.

TABLE 11. *Generalized linear mixed-effects model (family: Gaussian) fitted to the voice onset times, using 'Inexperienced–News report stimuli' as the reference condition*

	Estimate	Std. Error	*t*	99 % CI	
(Intercept)	0.476	0.017	28.70	*0.434, 0.520*	
Itemtype_JobAd	0.010	0.016	0.61	*–0.031, 0.048*	
Group_Recruiters	0.066	0.018	3.56	*0.016, 0.115*	**
Group_Jobseekers	0.064	0.018	3.53	*0.017, 0.111*	**
Itemtype_JobAd x Group_Recruiters	–0.030	0.007	–4.12	*–0.048, –0.011*	**
Itemtype_JobAd x Group_Jobseekers	–0.019	0.005	–3.43	*–0.033, –0.005*	**

NOTE: Significance code: 0.01 '**'.

of the target word in letters (WORDLENGTH), and its lemma-frequency, residualized against word length (rLOGFREQ), as they are known to affect naming times. In addition, we examined possible effects of PRESENTATIONORDER and BLOCK, as artifacts of our experimental design. Furthermore, we investigated three different operationalizations of word predictability. GENERICSURPRISAL is the surprisal of the target word given the cue, estimated by language models trained on the generic corpus

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

meant to reflect Dutch readers' overall experience. ClozeProbability amounts to the percentage of participants that complemented the cue with the target word in the Completion task preceding the VOT task. The binary variable TargetMentioned indicates whether or not the target word had been mentioned by a given participant in the Completion task. The fixed factors were standardized to reduce collinearity between predictors. Participants and items were included as random effects. We incorporated a random intercept for both items and participants to account for between-item and between-participant variation. We then added fixed effects one by one and assessed by means of likelihood ratio tests whether or not they significantly contributed to explaining variance in voice onset times.

We started with WordLength ($\chi^2(1) = 13.73$, $p < .001$), followed by rLogFreq ($\chi^2(1) = 4.78$, $p < .05$), and PresentationOrder ($\chi^2(1) = 3.97$, $p < .05$). After that, we added block ($\chi^2(1) = 2.10$, $p = .15$) and the interaction term PresentationOrder x block ($\chi^2(1) = 0.01$, $p = .93$). Given that neither of the latter two improved model fit, we left out these predictors. We then proceeded with the predictability measures, starting with the most general one: GenericSurprisal. This predictor did not contribute to the fit of the model ($\chi^2(1) = 2.54$, $p = .11$) and therefore we omitted it. ClozeProbability did improve model fit ($\chi^2(1) = 49.22$, $p < .001$), as did TargetMentioned ($\chi^2(1) = 309.37$, $p < .001$). We then included the interaction term rLogFreq x ClozeProbability, which did not contribute to the fit of the model fit ($\chi^2(1) = 3.60$, $p = .06$). rLogFreq x TargetMentioned did explain a significant portion of variance ($\chi^2(1) = 16.75$, $p < .001$). Finally, none of the two-way interactions of PresentationOrder and the other predictors in the model was found to improve model fit (PresentationOrder x TargetMentioned ($\chi^2(1) = 0.57$, $p = .45$); PresentationOrder x ClozeProbability ($\chi^2(1) = 0.65$, $p = .42$); PresentationOrder x rLogFreq ($\chi^2(1) = 0.21$, $p = .65$); PresentationOrder x WordLength ($\chi^2(1) = 2.58$, $p = .11$)). The model selection procedure thus resulted in a model comprising WordLength, rLogFreq, PresentationOrder, ClozeProbability, TargetMentioned, and rLogFreq x TargetMentioned.

We then added random slopes for participants. There are no by-item random slopes, because each item has only one lemma frequency, one cloze probability, one corpus-based surprisal estimate, one length, and a fixed position in the presentation order. Furthermore, there are items no one had mentioned in the Completion task, thus prohibiting by-item random slopes for TargetMentioned. Within these limits, a model with a full random

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press

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, Kliegl, Vasishth, Baayen, & Bates, 2017). Model comparisons indicated that the inclusion of the by-participant random slopes for WordLength, PresentationOrder, ClozeProbability, and TargetMentioned was justified by the data ($\chi 2(5) = 53.00$, $p < .001$). Then, confidence intervals were estimated via parametric bootstrapping over 1,000 iterations (Bates et al., 2015). We first ran the model using 'Target not mentioned' as the reference condition and then 'Target mentioned'. The outcomes are presented in Table 5 in Section 4.2.

https://doi.org/10.1017/langcog.2018.4 Published online by Cambridge University Press