Cognition 125 (2012) 80–106

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Probability and surprisal in auditory comprehension of morphologically complex words

Laura Winther Balling [a], R. Harald Baayen [b,*]

[a] Department of Aesthetics and Communication, University of Aarhus & Department of International Business Communication, Copenhagen Business School, Denmark

[b] Seminar für Sprachwissenschaft, Eberhard Karls Universität, Tübingen & Department of Linguistics, University of Alberta, Edmonton, Canada

ARTICLE INFO

Article history:
Received 26 February 2011
Revised 16 April 2012
Accepted 10 June 2012
Available online 27 July 2012

Keywords:
Spoken word recognition
Shortlist B
Morphological processing
Uniqueness points
Neighborhood measures
(cumulative) Surprisal
Kullback–Leibler divergence
Morphological family size

ABSTRACT

Two auditory lexical decision experiments document for morphologically complex words two points at which the probability of a target word given the evidence shifts dramatically. The first point is reached when morphologically unrelated competitors are no longer compatible with the evidence. Adapting terminology from Marslen-Wilson (1984), we refer to this as the word's initial uniqueness point (UP1). The second point is the complex uniqueness point (CUP) introduced by Balling and Baayen (2008), at which morphologically related competitors become incompatible with the input. Later initial as well as complex uniqueness points predict longer response latencies. We argue that the effects of these uniqueness points arise due to the large surprisal (Levy, 2008) carried by the phonemes at these uniqueness points, and provide independent evidence that how cumulative surprisal builds up in the course of the word co-determines response latencies. The presence of effects of surprisal, both at the initial uniqueness point of complex words, and cumulatively throughout the word, challenges the Shortlist B model of Norris and McQueen (2008), and suggests that a Bayesian approach to auditory comprehension requires complementation from information theory in order to do justice to the cognitive cost of updating probability distributions over lexical candidates.

1. Introduction

The Shortlist B model proposed by Norris and McQueen (2008) is the most comprehensive computational theory of auditory comprehension available to date. The model computes that sequence of words that is most likely to represent the lexical parse of the utterance heard, given the input, a stream of phonemes coming in over time. For instance, given the Dutch sequence of spoken words *kar personen*, the final state of the model is one in which *kar* ('cart') and *personen* ('persons') have probability 1, whereas competitors such as *karper* ('carp') and *persoon* ('persons') have probability 0.

This example (see Norris & McQueen, 2008, p. 370) illustrates that Shortlist B is a full listing model in the sense of Butterworth (1983) and Janssen, Bi, and Caramazza (2008), in that its lexicon contains entries for morphologically complex words such as *personen*. In the present example, the plural form *personen* suppresses its singular *persoon* as soon as the evidence for the plural suffix becomes available. The characterization of Shortlist B as a full form model for morphological processing is supported by an examination of its lexicon. The Shortlist B simulations reported by Norris and McQueen (2008) are all based on a lexicon of 20,250 Dutch word forms. This set of words combines two subsets: the 20,000 most frequent word forms in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), as well as an additional 250 words not

* Corresponding author. Tel.: +49 (0) 7071 29-73117; fax: +49 (0) 7071 29-5818.
E-mail addresses: baayen@ualberta.ca, harald.baayen@gmail.com (R.H. Baayen).

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.06.003

in this list, but required for the simulation of experimental data. A perusal of the 20,000 most frequent word forms in CELEX indicates that at least half of these word forms are derived or compound words. Furthermore, of those words that are not compounds or derived words, roughly a third carry an inflectional ending. The full-form approach of Shortlist B to morphological processing sets this model apart from theories assuming obligatory decomposition (see, e.g., Rastle & Davis, 2008, for visual comprehension and Marslen-Wilson, Tyler, Waksler, & Older, 1994, for auditory comprehension).

There are two potential reasons why it is advantageous for Shortlist B to ignore a word's internal structure. One is that including morphological structure would lead to a computationally much more complex model, with additional layers of Bayesian inference for the probability of a complex word given its morphemic constituents. To illustrate the kind of complexities that would arise, consider the consequences of the Possible Word Constraint (PWC; Norris, McQueen, Cutler, & Butterfield, 1997) for inflected words. The PWC states that a word candidate is disfavored in recognition of continuous speech if accepting that word candidate means that a string which does not represent a valid word in the language is left over. In English, the PWC would penalize words such as *sea* in the string *seash* because *sh* is not a possible word of English. In Dutch, one-phoneme inflectional suffixes such as the -*t*/-*d* and -*s* suffixes would complicate the implementation of the PWC. Although the PWC would work well for a word such as Dutch *mat* ('mat'), correctly penalizing *ma* ('mom'), it would run into problems for *bakt* ('bakes'). The stem *bak* would be penalized, even though it is a legitimate and well-formed part of the inflected form *bakt*. Implementing an exception status to the possible word constraint for the -*t*-suffix may cause as many problems as it solves. More generally, solving the problem of segmenting the stream of phonemes into a non-overlapping sequence of words becomes considerably more complex in a full decomposition approach. Moreover, a full decomposition representation may result in a large number of spurious morphological parses during word recognition, such as parsing the monomorphemic *mat* into the non-constituents *ma* and -*t* (Baayen & Schreuder, 2000).

Second, recognizing a word amounts to accessing its meaning. When in a visual world paradigm, participants shift their gaze from a picture of a ham to a picture of a hamster when listening to the word *hamster*, this is because they have understood that the more likely meaning conveyed by the incoming speech signal is not 'ham' but 'hamster' (cf., e.g., Salverda, Dahan, & McQueen, 2003). Since for derived words and compounds, the complex word often carries shades of meaning that are not straightforwardly predictable from the meanings of its parts, the proper interpretation of such words is achieved by having independent representations for complex words, each associated with its own a priori likelihood and its own specific phoneme sequence. As the speech signal unfolds, the constituents of a complex word may develop higher activation levels, but eventually they have to give way to the complex word, which has more bottom-up support. Thus, Shortlist B is compatible with parallel dual route models such as those proposed by Baayen, Dijkstra, and Schreuder (1997), Baayen and Schreuder (1999) and Baayen, McQueen, Dijkstra, and Schreuder (2003), in which full forms and their constituents are in competition. Crucially, in all these models it is not the case that a complex word can be understood only after its constituents have been accessed, as in obligatory decomposition models.

The aim of the present study is to clarify whether Shortlist B correctly predicts the processing costs of morphologically complex words presented in isolation to the listener. Shortlist B predicts for complex words that lexical access is completed when a lexical representation fully covering the evidence in the input has reached a posterior probability close to 1. Just as the sequence of words *kar personen* is resolved at the moment that a morphologically unrelated competitor such as *karper* and a morphologically related competitor such as *persoon* are suppressed, a compound such as *blackboard* can only reach threshold probability by suppressing unrelated competitors such as *boar* and *lack* and constituents such as *black* and *board*. The point in time at which the whole word has suppressed its competitors and has attained a posterior probability close to unity will be the point at which a lexical decision can be initiated. Hence, Shortlist B predicts the uniqueness point of the whole word to be a predictor for response latencies.

Interestingly, Shortlist B predicts that the uniqueness point of the first constituent of a complex word should not correlate with response latencies. Shortlist B allows competitors such as *karper* in *kar personen* to reach a probability of 1 before being suppressed by the correct parse of the input, *kar* and *personen*. Similarly, for *blackboard*, *black* will first be a certain candidate, before being downgraded by *blackboard*. Lexical decisions cannot be based on such early highly activated competitors, which may or may not be morphologically related. In the Bayesian framework, therefore, the time at which a sufficiently high posterior probability for the full input stream is obtained is the crucial predictor for response latencies. Points of disambiguation upstream are irrelevant.

In what follows, we present two auditory lexical decision experiments that demonstrate that there is an effect on response latencies of a UP before the whole-word UP, contrary to the predictions of Shortlist B. A third, visual lexical decision experiment will then demonstrate that this early UP is so strong that it is even predictive for visual lexical decision latencies. To understand this early UP effect, following up on work on sequence processing in syntax (Hale, 2001; Levy, 2008), we will make use of information theory, and specifically the measures of Kullback–Leibler divergence and surprisal, as estimators of cognitive processing costs associated with updating Bayesian probability distributions calculated over a full-form lexicon.

2. Experiment 1

Experiment 1 addresses the auditory comprehension of complex words in Danish, focusing on two critical points in the resolution of lexical competition, to which we henceforth refer as UP1 (for initial uniqueness point) and CUP (complex uniqueness point). Whereas an effect of CUP is compatible with Shortlist B, an effect of UP1 is not. Before discussing the technical details of Experiment 1, we first introduce these two uniqueness points.

2.1. Uniqueness points

The standard UP is defined as the point at which a word deviates from all onset-aligned words in the language excepting suffixed words and compounds (Marslen-Wilson, 1984, Marslen-Wilson & Welsh, 1978). Without this exception clause, the UP would occur after word offset for the majority of simple words. Thus, for example, the word-initial cohort for the English word *kind* includes *kin* and *kite*, but not *kindness* or *kindly*; the UP of *kind* occurs at the *n* where *kind* becomes distinguishable from *kite*. This formulation of the UP rules out the possibility that morphologically complex words, including those that are related to the target word, may play a role in the recognition process. Therefore, Wurm, Ernestus, Schreuder, and Baayen (2006) considered a new cohort-based measure, Shannon's entropy calculated across what we will refer to as the set of continuation forms: Words that are morphologically related continuations of the target word. For *kind*, for instance, the continuation set comprises the words *kind-hearted, kind-heartedly, kind-heartedness, kindliness, kindly*, and *kindness*. All these morphological relatives

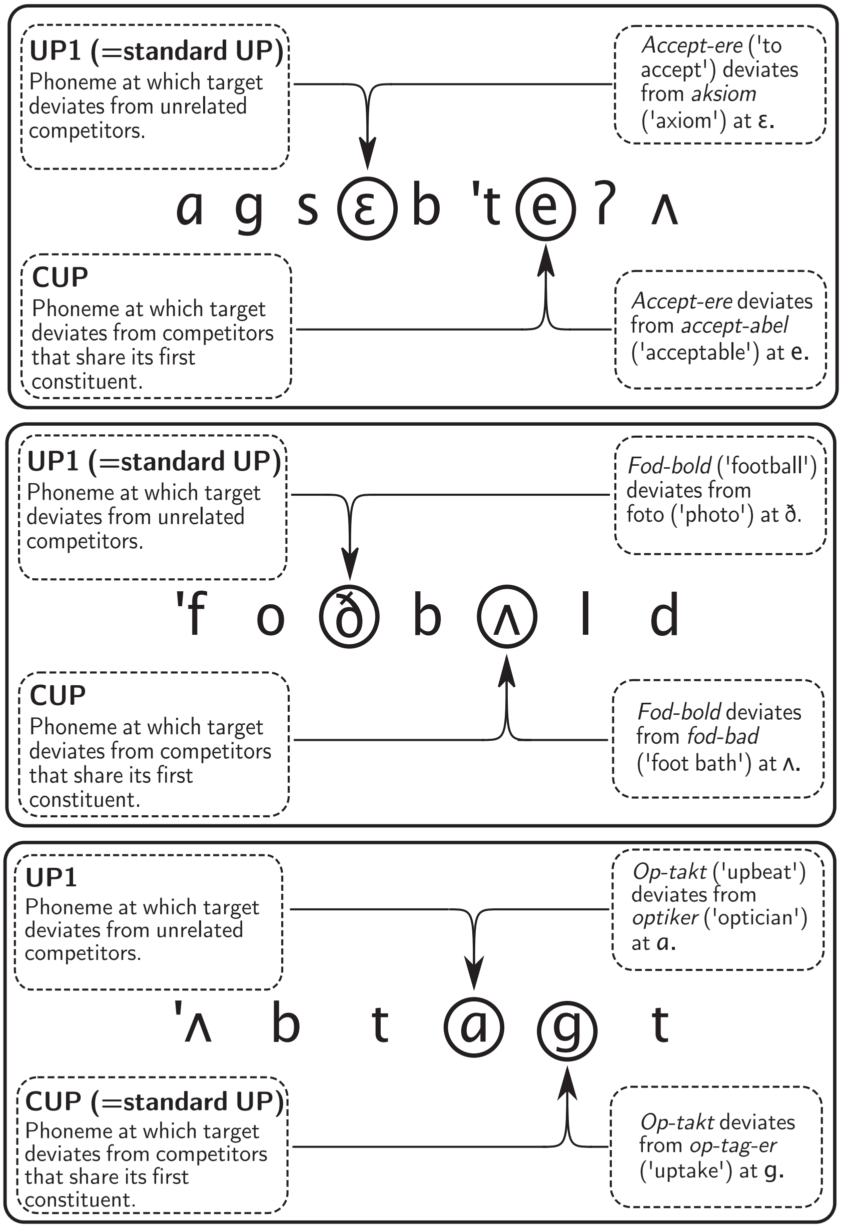
UP1 (=standard UP)
Phoneme at which target deviates from unrelated competitors.
Accept-ere ('to accept') deviates from aksiom ('axiom') at ɛ.
a g s ɛ b 't e ʔ ʌ
CUP
Phoneme at which target deviates from competitors that share its first constituent.
Accept-ere deviates from accept-abel ('acceptable') at e.
UP1 (=standard UP)
Phoneme at which target deviates from unrelated competitors.
Fod-bold ('football') deviates from foto ('photo') at ð.
'f o ð b ʌ l d
CUP
Phoneme at which target deviates from competitors that share its first constituent.
Fod-bold deviates from fod-bad ('foot bath') at ʌ.
UP1
Phoneme at which target deviates from unrelated competitors.
Op-takt ('upbeat') deviates from optiker ('optician') at ɑ.
'ʌ b t ɑ g t
CUP (=standard UP)
Phoneme at which target deviates from competitors that share its first constituent.
Op-takt deviates from op-tag-er ('uptake') at g.

Fig. 1. Definitions of UP1 and CUP and examples for Danish, for the suffixed word accept-ere ('to accept'), the compound *fod-bold* ('football'), and the (particle) prefixed word op-takt ('upbeat').

of *kind* are systematically ignored in the calculation of the standard UP. Nevertheless, Wurm et al. (2006) observed significant effects of Shannon's entropy calculated across the sets of continuation forms. The more continuation forms a word had and the more equal the probabilities of these continuations, the faster the word was responded to in auditory lexical decision.

More recently, Balling and Baayen (2008) introduced the Complex UP (henceforth CUP), a new UP specifically designed for gauging the role of morphologically related words in the post-UP cohort for morphologically complex target words. The CUP is the point at which a suffixed word becomes uniquely distinguishable from all words that share the same stem, with the exception of those words that are continuation forms of the suffixed target word itself. In the case of *kindly*, the CUP is at the *l*, where *kindly* diverges from *kindness* and *kind-hearted*. For the Danish suffixed word *accept-ere* ('to accept'), the UP occurs when *accept-ere* deviates from the unrelated word *aksiom* ('axiom'), and the CUP occurs when *accept-ere* deviates from the related word *accept-abel* ('acceptable'). This is shown in the top panel of Fig. 1 (with the standard UP termed UP1). Here, we disregard the continuation forms of *kindly* itself—*kindliness* is not taken into consideration—for the very same reason that continuation forms are discarded when calculating the standard UP. Balling and Baayen (2008) reported significant effects of both the standard UP and the new CUP in auditory lexical decision to suffixed words in Danish. The longer unrelated words were compatible with the target, the longer it took listeners to respond, as indexed by the significant effect of the standard UP. The duration of the competition from morphologically related words, indexed by the CUP, also revealed a significant effect on recognition, with longer latencies to words with later CUPs.

For compound words, the standard UP defines the position where a compound deviates from all other words in the language except continuations of its own first constituent. The UP, which we henceforth also refer to as UP1, thus indexes competition from morphologically unrelated words. The CUP, by contrast, occurs when the second constituent of a compound deviates from other second constituents attested as continuations of the first constituent. For a compound such as *kind-hearted*, the CUP is at the *h* where *kind-hearted* deviates from *kind-ness*. Similarly, the Danish compound *fod-bold* ('football') deviates from the last unrelated competitor *foto* ('photo') at the *d* and from the last related competitor, *fod-bad* ('foot bath'), at the second *o*, as illustrated in the middle panel of Fig. 1. The CUP can thus be used to index exactly the same kind of competition dynamics for both suffixed words and compounds.

For prefixed words, we continue with the same logic. As for suffixed words and compounds, we define the UP for a prefixed word as the point at which all morphologically unrelated competitors cease to be fully compatible with the input. For the prefixed Danish word *op-takt* ('upbeat'), this would be the *a* where *op-takt* deviates from *optiker* ('optician'), as illustrated in the lower panel of Fig. 1. We henceforth also refer to this uniqueness point as UP1. Note that the terminology of UP1 (as distinct from UP) is motivated by this uniqueness point, which differs from the traditional uniqueness point, which for prefixed words occurs later in the word: The classic UP does not distinguish between morphologically related and morphologically unrelated competitors for prefixed words. For instance, in *op-takt* the classical UP occurs at the *k*, where *op-takt* deviates from the related word *op-tag-er* ('to take up' or 'to record'). The UP1, however, is at the *a*, where the last morphologically unrelated words begin to mismatch the speech signal (e.g., *optiker*).

The CUP for prefixed words is defined as the point where the prefixed target deviates from the group of words that share the same prefix, i.e., for *op-takt*, the *k* where *op-takt* deviates from *op-tag-er*. The definition of the CUP is thus the same across all types of morphologically complex words, irrespective of whether the first constituent is a prefix, the stem of a suffixed word, or the first constituent of a compound; this is illustrated in Fig. 1. For both UP1 and CUP, continuation forms of the target words themselves are excluded from the computations. We will probe these continuation sets in our experiments by means of a separate measure, the cardinalities of these continuation sets. For words with two morphemes, such as the items in the present experiments, we thus predict effects of both UP1 and CUP, measured in milliseconds, while we use the number of continuations to probe the influence of the cohort at word offset. For words with more than two morphemes, we might in addition to UP1 and CUP effects observe competition beyond the CUP between candidates that share the first two morphemes. Such competition would be resolved at a second complex UP, the position of which may be predictive for recognition of multimorphemic words. However, at this point cohorts tend to be much reduced and any competition would be likely only to have small effects.

The CUP for prefixed words is very similar to the Conditional Root Uniqueness Point (CRUP) introduced by Wurm (1997), see also Wurm and Ross (2001) and Wurm et al. (2006). The CRUP is defined as the point at which the stem becomes distinguishable from all other free stems that can combine with the prefix heard, while the CUP is the point where the target becomes unique from all other words in the language that share the same prefix, whether the stem is free or bound. Wurm and collaborators showed that prefixed words for which the CRUP precedes the classical UP are processed faster than matched controls for which CRUP and UP coincide. However, for Danish, as compared to English, there are only very few words for which the last competitor for a prefixed word is a morphologically unrelated word. This happens for only two out of 175 prefixed words used as stimuli in the experiments reported below. Occasionally, unrelated and related competitors become incompatible with the input at the same phoneme; this is the case for 18 out of 175 prefixed items in the experiments reported below. We therefore were not able to explore experimentally for Danish the potential advantage for a prefixed word of having the CRUP preceding the classic UP, but instead probe the time course of competition using the UP1 and CUP measures for prefixed as well as for other types of complex words.

2.2. Method

2.2.1. Materials

We selected 150 bimorphemic Danish words, mostly nouns and verbs, but also some adjectives, for presentation. These are listed in the appendix, along with glosses and translations. The experiment also included 50 simple word fillers and 200 nonwords. As detailed below, the bimorphemic words comprised prefixed words, particle prefixed words, and compounds. Candidate items were randomly selected from the Danish vocabulary, as represented in a corpus of 43.6 million Danish words (for details, see Balling, 2008, chapter 3), and items were then selected that fulfilled a range of criteria outlined in this section. By accident, two of the complex words chosen contained allomorphs of the same stem. Both were presented in the experiment, always in the same order. Only the first word presented was included in the analyses.

All stems of prefixed and particle prefixed words and both constituents of the compounds were morphemes that can also be used as independent words. Words with linking elements and stem allomorphy were avoided, with the exception of regular stress- and *stød* (glottal stop)-variations and schwa deletion. Highly irregular pronunciations were also avoided, as were pronunciations varying substantially between casual and careful speech. We avoided homonymous or strongly polysemous words. Homonymous constituents could not be avoided entirely; however, we made sure that all constituents were unambiguous in their target words. We likewise avoided semantically opaque complex words. For the compounds, we considered the transparency of both constituents in relation to the meaning of the compound (cf. Libben, Gibson, Yoon, & Sandra, 2003). Many compound verbs in Danish carry one of a relatively restricted number of verbs with rather broad and vague meanings. Only two such compounds were included.

The 200 nonwords were constructed by changing one to three phonemes in the stems of the real words. All prefixes and particle prefixes were retained on the nonwords, to avoid that the presence of an affix alone could be enough to make a word decision. For compounds, both constituents were changed into nonwords, as pretesting indicated that including real stems in nonce compounds would make such compounds disproportionally difficult to reject.

Thirty words and nonwords were used for training and warm-up. These items had a similar composition to the experimental items, but carried different affixes in order not to introduce variations in the number of times each of the experimental affixes was encountered.

The stimuli were recorded in a quiet room by a female native speaker of Danish directly onto a hard disk at a sampling rate of 48 kHz and a bit depth of 16 bit. Words and nonwords were mixed in the reading lists, with reading fillers at the beginning and end of the lists in order to avoid beginning- and end-of-list intonation on the items. The items were normalized for peak intensity.

2.2.2. Predictors

The critical variables in this experiment are the UP and CUP measures. We considered these measures jointly with a range of other measures in a regression design. Table 1 lists these measures, together with their mean, standard deviation, and range.

2.2.2.1. UP1 and CUP. The two central UP measures were determined in the following way. We queried the Danish corpus for a beginning-of-string marker combined with possible spellings of increasingly larger parts of a phonological transcription of the given target word. UP1 was defined as the position at which the query returned only words that were morphologically related continuation forms of the first constituent. CUP was defined as the position at which the query returned only continuation forms of the whole word. The phonemes carrying the uniqueness points were then located in the speech signal, based on waveforms and spectrograms in the waveform editor Cool Edit 2000. The uniqueness point was defined as the middle of the time segment corresponding to the relevant phoneme, except that for stop sounds the beginning of the release noise was defined as the uniqueness point. Uniqueness points in milliseconds were then defined as the duration of the signal from word onset to these locations.

The uniqueness point measures and word duration are highly collinear: They are all durations measured from word onset. To reduce collinearity (which is problematic for the regression analysis), we partitioned the auditory signal into three non-overlapping parts: the distance from word onset to UP1, the distance from UP1 to CUP, and the distance from CUP to word offset.

2.2.2.2. Word type. The bimorphemic words comprised 50 compounds, 50 prefixed derived words, and 50 words car-

Table 1
Lexical predictors for the items in Experiment 1 The variables marked with an asterisk are frequency counts per million words based on a 43.6 million word corpus of Danish. All other variables are based on the same corpus. $N = 149$.

Predictor	Mean	SD	Range
UP1, ms	297	83	141–847
Complex UP, ms	499	112	290–847
Duration, ms	747	122	450–1210
Length in phonemes	7.3	1.6	4–15
Continuation forms (type frequency)	17	78	0–924
Cohort density	383118	301760	318–1338611
Neighborhood size	0.4	0.8	0–6
Whole-word frequency*	4	17	0–186
Second constituent frequency*	401	2658	0.2–32437
Family size, first constituent	304	354	5–1476
Family size, second constituent	611	606	4–3476
Mean bigram frequency*	26124	13991	1445–70722
Juncture bigram frequency*	10033	16712	23–83663

rying particle prefixes. Particle prefixes are formatives that can function both as prefixes and as independent prepositions or particles. Some verbs can occur separated from their particle (as in Dutch and German), for other verbs this is impossible; we included only the latter kind of verbs. For nouns and adjectives carrying particle prefixes, such separation is never possible. Prefixation, particle prefixation, and compounding are all productive in Danish, with compounding responsible for the majority of new words (Hansen, 1967, p. 241). As a consequence of the high degree of productivity of compounding, morphological families in Danish tend to be much larger than those of English or Dutch (see Balling, 2008, pp. 85–88). In what follows, word type denotes the factor distinguishing between compounds, prefixed words, and particle prefixed words.

The particle prefixed words carried five different particles, each represented by ten words. These particles were comparable to the prefixes in length. Likewise, five derivational prefixes were used, each of which also occurred on ten different target words. The affixes used in the experiment, and translations of them, can be found in the appendix.

2.2.2.3. Cohort density. We also included as a predictor the cohort density measure proposed by Magnuson, Dixon, Tanenhaus, and Aslin (2007): the summed log frequency of the words that overlap with the target on the first two phonemes. Like the UPs, this measure indexes similarity from word onset between the target and the rest of the vocabulary, but it is based on frequencies of competitors rather than the time-course of disambiguation. We also examined a measure of global similarity across all phoneme positions, using the N-count neighborhood density count. This predictor never reached significance in our analyses. We used our written corpus of Danish as an index of both phonological cohorts and neighborhoods, in the absence of sufficiently large phonologically transcribed Danish corpora.

2.2.2.4. Morphological family size. A morphological factor that has been documented to play a role especially in visual word recognition is morphological family size: the type frequency of the derivations and compounds that share the stem of a target word. Words with larger families tend to be easier to recognize (Moscoso del Prado Martín et al., 2005; Schreuder & Baayen, 1997). For auditory word recognition, however, Baayen, Wurm, and Aycock (2007) report no effects of morphological family size measures in the auditory modality at all, while Meunier and Segui (1999) observed an inhibitory effect instead of a facilitatory effect, and only for family members of higher frequency than the target. Our hypothesis is that, since the auditory signal unfolds over time, mainly onset-aligned family members are relevant in auditory processing.

First and second constituent family sizes were extracted from the Danish corpus. The first constituent family size counts were restricted to those family members for which the shared constituent also occurred in the first position (De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002). For words with prefixes and particle prefixes, morphological families were further restricted to those complex words in which the prefix occurred in the outermost layer of its derivational structure. For second constituents, the family size counts were not position-specific. Extracting position-specific family counts by hand turned out to be undoable given a corpus without morphological mark-up.

2.2.2.5. Continuation count. The count of family members that are onset aligned with the target, and contain the target as a constituent.

2.2.2.6. Frequency. We considered several frequency measures as predictors. Whole-word frequency was defined as the string frequency of the complex form as it was presented in the experiment. We also examined constituent frequency measures. The second constituent frequency was defined as the lemma frequency of the second constituent, i.e., the summed frequency of all inflectional variants of that constituent. To anticipate the results, second constituent frequency did not emerge as a significant predictor. We also considered the surface form frequency of the second constituent instead of its lemma frequency, but this predictor did not fare any better. Similarly, we did not observe significant effects of first constituent frequency for compounds and particle prefixed words for which this measure was appropriate.

2.2.2.7. Bigram frequency. As further controls, we considered two bigram measures, the mean bigram frequency of all letter pairs in the word and the frequency of the bigram straddling the morpheme boundary. The mean bigram frequency provides some control of unusual phoneme sequences. The juncture bigram measure was included in order to gauge whether low-frequency transitions might favor morphological decomposition (Cutler, 1981; Hay, 2002; Seidenberg, 1987) or strengthen hypotheses about potential word boundaries (Norris, 1994).

2.2.2.8. ISI. We manipulated the Inter-Stimulus Interval (ISI) in order to investigate whether effects would vary systematically with the pace of the experiment, with a fixed ISI of 3000 ms resulting in a slower paced and a variable ISI resulting in a faster paced experiment.

2.2.2.9. PC1–PC4. Reaction times may enter into strong correlations with reaction times at previous trials: The response latencies of a given participant often constitute a time series in which the response at trial t is correlated with the responses at preceding trials $t-1, t-2$, etc. Following Baayen et al. (2007), De Vaan, Schreuder, and Baayen (2007), and Baayen and Milin (2010), we sought to bring at least some of these cross-trial dependencies under statistical control. We restricted ourselves to the four preceding trials. As the response latencies at these trials are highly collinear, we orthogonalized them using Principal Components Analysis, resulting in four principal components, henceforth PC1–PC4. In addition to the latencies on previous trials, we considered the lexicality of the previous item, the correctness of the previous response, and the trial number as predictors that were included in order to control the effects of experimental context.

2.2.3. Participants

40 volunteers were tested individually in a sound-attenuated room. There were 12 males and 28 females, between the ages of 21 and 41 (mean 29.5). All participants had grown up with Danish as their first language and reported normal hearing. Most were students at the University of Aarhus.

2.2.4. Procedure

The experiment was run on a portable computer, using DMDX (Forster & Forster, 2006). Stimuli were presented over headphones. Participants received standard lexical decision instructions in writing and were allowed to ask questions after a practice session consisting of 20 items. Six warm-up items were presented at the beginning of the experiment and two warm-up items after each of the two breaks which occurred one third and two thirds through the experiment.

Each trial began with a fixation point (a plus) displayed in the middle of the screen for 500 ms after which a stimulus was played. For half the participants, ISI was variable: a trial ended when the participant responded or at a time-out of 3000 ms from the beginning of the trial. For the other half of the participants, ISI was fixed at 3000 ms. When ISI was variable, the experiment lasted 10–15 min; when it was fixed, the experiment lasted approximately 25 min. Participants indicated their lexical decision by a button press, using their dominant hand to indicate a yes decision.

Each participant heard a different pseudo-random order of the stimuli. No more than three words or nonwords occurred in a row and no prefix appeared on consecutive trials. The stimulus orders were generated using Mix (Van Casteren, 2006).

2.3. Results and discussion

For the analysis of the response latencies, we excluded error responses (3.9%) as well as all responses to two very low-frequency prefixed words with error rates exceeding 30%, leading to a total data loss of 4.4%. All data points were retained in the analysis of accuracy. Response latencies as well as UPs, frequency and family size measures were logarithmically transformed to reduce the likelihood of atypical outliers dominating the analyses. For variables for which the lowest values were zero, we added 1 to the value before carrying out the transformation.

We used generalized additive linear mixed-effects regression models for the analyses (Baayen, Davidson, & Bates, 2008; Wood, 2006) relying on the R environment for statistical computing (R Development Core Team, 2011) and the mgcv package (Wood, 2006, 2011). The models included random intercepts for participant and item. These random intercepts were supported by likelihood ratio tests with p-values below 0.05; exploratory analyses showed that no other random effects (such as affix) were justified. Random slope parameters, such as for instance random slopes for word frequency by subject, were tested for, but excluded from the final model when not supported by likelihood ratio tests (i.e. $p > 0.05$). Specifically, we tested whether random slopes for the central UP-variables were justified, which was not the case. The models summarized in Tables 2–4 were reached by initially fitting models to the latencies and binary choice data using all potentially relevant predictors that we had available to us. We then removed predictors that were non-significant at the 0.05 level step by step (but keeping in the model non-significant effects for predictors that participated in significant higher-order interactions). Inspection

Table 2
Parametric coefficients of the generalized additive mixed model fitted to the response latencies of Experiment 1.

	Estimate	Standard error	*t*	*p*
Intercept	6.9061	0.0278	248.3007	0.0000
Previous RT PC1	0.1089	0.0076	14.3431	0.0000
Previous RT PC3	0.0294	0.0091	3.2489	0.0012
Previous RT PC4	0.0426	0.0092	4.6257	0.0000
ISI:Fixed	0.0415	0.0369	1.1242	0.2610
Previous Item:Word	0.0344	0.0046	7.4713	0.0000
Trial	−0.0001	0.0000	−4.3920	0.0000
Log Word Frequency	−0.0111	0.0028	−3.9731	0.0001
Residualized Continuation Forms	−0.0129	0.0056	−2.3087	0.0210
ISI:Fixed * Previous Item:Word	−0.0271	0.0065	−4.1880	0.0000
ISI:Fixed * Trial	0.0001	0.0000	2.7077	0.0068
ISI:Fixed * Log Word Frequency	−0.0074	0.0018	−4.1270	0.0000

Table 3
Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), *F* and *p* values for the splines and random effects in the generalized additive mixed model fitted to the response latencies of Experiment 1.

	edf	Ref.df	*F*	*p*
spline Log UP1	2.6649	2.7502	66.3970	0.0000
spline Log UP1toCUP	3.4700	3.5786	37.7404	0.0000
spline Log CUPtoOffset	3.7766	3.8926	16.4042	0.0000
random intercepts Word	107.8014	120.9114	5.5453	0.0000
random intercepts Participant	37.6941	37.9974	36.5426	0.0000

Table 4
Parametric coefficients of the generalized additive mixed model fitted to the response correctness of Experiment 1. No smoothers or random effects reached significance.

	Estimate	Standard error	*z*	*p*
Intercept	−0.2802	0.5162	−0.5428	0.5873
Type:Prefix	1.1901	0.1928	6.1740	0.0000
Type:Particle	0.0663	0.2074	0.3196	0.7493
Log Word Frequency	−0.7634	0.0593	−12.8745	0.0000
Residualized Continuation Forms	−0.9363	0.1163	−8.0505	0.0000
Log Juncture Bigram Frequency	−0.1551	0.0417	−3.7201	0.0002

of the distribution of the residuals of the resulting model for the response latencies revealed a marked departure from normality. We therefore removed potentially overly influential outliers (2.8% of the responses, characterized by standardized residuals exceeding −2.5 or +2.5) and refitted the model (Crawley, 2002; Baayen & Milin, 2010). Predictors that did not reach significance, and that were therefore removed in a stepwise variable elimination procedure, are not listed in the tables of coefficients reported here, or for Experiments 2 and 3 below.

Table 2 presents the parametric coefficients of the generalized additive mixed model for the response latencies, while Table 3 shows the non-linear terms and random effects of that model. Table 4 shows the generalized additive model fitted to response correctness for Experiment 1.

As mentioned above, the UP measures were decorrelated by making them index non-overlapping parts of the word. For 11 words, this recalculation resulted in zero values on either UP1 to CUP or CUP to offset. These 11 words were excluded from the models reported in Tables 2–4 to ensure that none of the effects were driven by these outliers. We also decorrelated whole-word frequency and the type count of morphologically related continuation forms by replacing the continuations variable with the residuals of a regression model with continuations as a function of whole-word frequency. This decorrelated variable is well correlated with the original continuations variable ($r = 0.80$), and hence can be straightforwardly understood as the number of continuations in so far as this cannot be predicted from whole-word frequency. The collinearity between the resulting set of predictors was thereby reduced to an acceptable level ($\kappa = 13.7$).

2.3.1. Control predictors

PC1, PC3 and PC4 were all significant predictors of the latencies. PC1, the principal component capturing most of the variance in the preceding reaction times (48%), had the largest effect size, spanning a range of latencies from arond 800 to around 1150 ms, a span of some 350 ms. The importance of taking this experimental ‘noise’ out of the error term may be appreciated by a comparison with the frequency effects in this experiment which had a span of just over 300 ms. None of the PCs entered into interactions with any of the other predictors. These principal component predictors indicate that long RTs on previous trials correlated with a long RT on the current trial.

In the variable-ISI version of the experiment, participants tended to respond slightly more quickly as they proceeded through the experiment. This effect of habituation, however, was absent in the slower version with the fixed ISI, as indicated by the interaction between ISI and Trial.

We also observed an interaction between ISI and the lexicality of the previous trial: whether the previous item was a word or nonword only played a role in the faster experiment with variable ISI; here, the responses were significantly longer when the previous item was a word than a nonword.

The number of times an affix had been repeated across the experiment had no significant effect. In other words, there was no measurable effect of within-experiment affix priming.

In the error analysis, none of these control predictors reached significance, unsurprisingly given the high level of accuracy.

2.3.2. Uniqueness points

There were significant, non-linear, positively accelerating effects of all three temporal measures: duration from word onset to UP1, from UP1 to CUP, and from CUP to word offset. These effects are illustrated in Fig. 2. In a regression model fitted to the data with the original UP measures, i.e. milliseconds from word onset to UP1 and CUP, both UP1 and CUP emerged with significant inhibitory effects, indicating that the results do not depend on how collinearity is reduced.

Complex words with a later UP1, as well as such words with a later CUP, elicited longer response latencies. While the effect of the CUP is predicted by Shortlist B, given that at the CUP a word’s posterior probability would be close to 1, the large and significant effect of UP1 challenges Shortlist B.

It is possible, however, that a frequency-weighted measure of the onset cohort may be a more precise measure of onset-based similarity than the present UP measures. We therefore included cohort density, the summed log frequency of the words overlapping with the target on the first two letters, in our analysis. This measure had no effect on response latency, and the UP1 and CUP measures remained significant when this non-significant predictor was included in the model specification. We tested whether making the high-frequency members of the onset cohort carry more weight, by using log summed frequency instead of Magnuson et al. (2007)’s summed log frequency, would make the effect significant; this was not the case. We also examined whether neighborhood density might explain the UP effects. However, there was no evidence whatsoever supporting neighborhood density as a predictor for the response latencies.

None of these measures interacted with ISI, indicating that task demands are not driving the effects observed. Furthermore, the effects of UP1, UP1 to CUP, and CUP to offset,

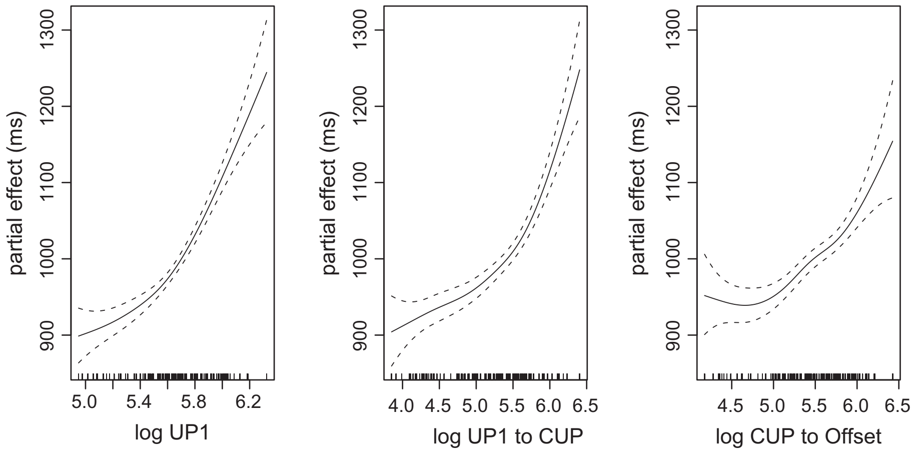

Fig. 2. Partial effects (shifted vertically by the intercept) of log UP1, log UP1 to CUP, and log CUP to Offset, on the RT in ms scale, using restricted cubic splines.

were identical for compounds, prefixed words, and words with particle prefixes. This suggests that the UP1 measure (which corresponds to the traditional UP for compounds but not for the two types of prefixed words) provides a uniform measure of competition from morphologically unrelated words.

2.3.3. Frequency measures

Consistent with the importance of whole-word frequency in Shortlist B, a significant effect of whole-word frequency was observed. By contrast, there were no significant effects of constituent frequencies, neither in the analysis of the response latencies nor in the analysis of the accuracy measure.

The effect of whole-word frequency was significantly facilitatory for both fixed and variable ISI, but the effect was stronger when ISI was fixed. This result could indicate that a slower pace allows better activation of lexical memory representations. The interaction could also be interpreted as evidence that the frequency effects observed in lexical decision tasks are partly caused by decision rather than recognition processes (Balota & Chumbley, 1984), such that longer decision times result in larger frequency effects. No other non-control variables interacted with ISI, suggesting that task demands play a relatively minor role in the pattern of results.

The error analysis also showed a significant effect of whole-word frequency, with a greater word frequency decreasing the likelihood of an error. The frequency of the bigram straddling the juncture between the constituent morphemes of the complex words also reached significance in the error analysis. Words with more frequent juncture bigrams were slightly less error-prone than those with a lower bigram frequency. No other effects of letter bigram frequencies were found.

2.3.4. Family size measures

The count of family members that left-embed the target word, and that are compatible with the target word up to target offset, had a small but significant facilitatory effect: The more continuation forms a word had, the faster it was responded to. Recall that we decorrelated the count of continuations from whole-word frequency. The analysis remained stable also if the direction of decorrelation was reversed so that whole-word frequency was decorrelated from continuations. Interestingly, this was the only cohort measure that also manifested itself in the accuracy analysis: Accuracy increased for words with more continuation forms, providing further evidence for the relevance of this measure for understanding auditory comprehension.

In this experiment, neither the classical family size count nor the related family frequency measure (the summed frequency of family members) emerged as significant predictors.

Family counts, as defined by Schreuder and Baayen (1997), comprise morphologically related forms, irrespective of whether they are onset-aligned with the target word. However, due to the dynamics of a system with continuously updated lexical probabilities as more bottom-up information becomes available over time, it is less likely that a compound such as *heartfelt* will become a high-probability lexical candidate competing with the target *kind-hearted* in auditory word recognition than it is in visual word recognition.

However, in the Shortlist B model, onset alignment is not a precondition for being a lexical competitor. As in the original Shortlist model (Norris, 1994), words can enter the competition at other points in the signal, for instance, when onset-aligned with embedded metrically strong syllables. Although words such as *heartfelt* may in principle come into play when *kind-hearted* is heard, the mismatch with the acoustic signal at the offset of *heart* may render competition from such words relatively harmless. In short, straightforward family counts probably are too coarse, and hence irrelevant for understanding auditory comprehension.

Interestingly, it is a subset of a word's morphological family, namely those family members that are onset-aligned with the target, that drives the effect of the CUP. While in visual lexical decision and reading, a greater family size leads to shorter response latencies and shorter fixation durations (De Jong, Schreuder, & Baayen, 2000; Kuperman, Schreuder, Bertram, & Baayen, 2009), in auditory lexical decision onset-aligned family members sharing the first (but not the second) morpheme have an inhibitory effect. It is only those family members that are fully consistent with all bottom-up information at word offset, the continuation forms, that show the facilitation familiar from the reading literature. This reversal from inhibition to facilitation for the continuation forms must be driven by the absence of substantive mismatching bottom-up information: The continuation forms fully match all segments of the target, but are longer.

It is noteworthy that the facilitatory effect of the number of continuation forms is not straightforwardly predictable from Shortlist B. At word offset, the continuation forms become incompatible with the evidence, hence their probabilities should go to zero. What we see, however, seems to indicate that their probabilities are merged with the target's probability. This merging might be driven by the semantic similarity shared by, for instance, a compound in the singular and its plural continuation form.

3. Surprisal in auditory comprehension

Experiment 1 provides unambiguous evidence that response latencies in auditory lexical decision to words with later UP1 and/or later CUP are longer compared to words with earlier UP1 and/or CUP.

Within the theoretical framework provided by the Shortlist B model, an effect of UP1 for *simple words* receives a straightforward interpretation. For simple words, UP1 indicates the point at which the probability mass of competing words has become negligible, allowing the target word to reach a critical threshold probability, on the basis of which a lexical decision response can be initiated. The earlier this critical point is reached, the earlier the response can be initiated.

For complex words, an effect of UP1 is not expected in the Shortlist B framework. In Shortlist B, a response latency hinges on, first, the target word covering the full input, and second, the target reaching a critical probability threshold close to one. (Recall that Shortlist B is a full-listing model in which complex words have to suppress their constituents.) These two conditions are not met at UP1. It is only at the CUP that the posterior probability of a complex target word will approach or cross the probability threshold for a response, while at the same time providing a complete covering of the input. Since response latencies are fully determined by this point in time, Shortlist B predicts that for complex words the CUP should be predictive, while the UP1 should be irrelevant.

To understand why such a large effect of UP1 is nevertheless present in our data, contradicting Shortlist B, we have to enrich Shortlist B with insights from information theory. As our point of departure, we take the surprisal theory of Hale (2001), Genzel and Charniak (2002), Genzel and Charniak (2003) and Levy (2008) for syntactic processing. Levy's central hypothesis is that the updating of the probability distribution of competing parses as a new word becomes available in the input constitutes an important cognitive bottleneck in sentence processing. Our hypothesis is that in exactly the same way, the updating of the probability distribution of competing words as a new segment comes in also constitutes a cognitive bottleneck. We propose that the underlying probabilistic mechanisms in syntactic processing and those in phonological processing are fundamentally the same.

Table 5
Example universe of strings and their a priori probabilities.

String	Frequency	Probability
abc	10	0.038
abd	40	0.151
abde	20	0.075
abdef	80	0.302
a	10	0.038
ab	5	0.019
zx	100	0.377

By way of illustration, consider Table 5, which lists a series of strings and their frequency. These strings can be understood either as words (sequences of phonemes), or as sentences (sequences of words). The relative frequencies of these strings constitute their a priori probabilities.

Consider the situation in which the string “ab” has been processed, and that the next element (word/segment) in the input to become available is “d”. Before the “d” comes in, the following strings are in the race, with the probabilities indicated.

```
abc        abd        abde       abdef      ab
0.06451613 0.25806452 0.12903226 0.51612903 0.03225806
```

We refer to this probability distribution as *Q*. After the “d” element comes in, the strings “ab” and “abc” become incompatible, and their probability mass is redistributed over the probabilities of the strings “abd”, “abde” and ”abdef”, resulting in a probability distribution *P*:

```
abd       abde      abdef
0.2857143 0.1428571 0.5714286
```

Focusing on the strings in *Q* with non-zero probability in *P*, we obtain the probability distribution Q':

```
abd       abde      abdef
0.2580645 0.1290323 0.5161290
```

By comparing *P* and Q', it is easily seen that the updated probability distribution *P* has higher probabilities than the ‘preceding’ probability distribution Q'. The difference between these two probability distributions can be assessed using the Kullback–Leibler divergence or relative entropy measure

$$\mathrm{RE}(P, Q') = \sum_i P_i \log_2 \frac{P_i}{Q'_i}. \tag{1}$$

For the present example, the relative entropy evaluates to 0.1468. This relative entropy is exactly the same as the information in bits of element “d” given the preceding string “ab”:

$$-\log_2 \frac{40 + 20 + 80}{40 + 20 + 80 + 10 + 5} = 0.1468. \tag{2}$$

This amount of information is known as the surprisal of element “d”. For a formal proof for the general case, the reader is referred to Levy (2008).

Levy (2008) and related work (see, e.g., Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Frank, 2009; Staub, 2010; Staub & Clifton, 2006) show that for syntactic processing a word with a high surprisal incurs a large processing cost, as indicated in the eye tracking record by measures such as fixation duration and number of regressions. Our hypothesis is that the same processing principles that lead to a cost of surprisal in syntax likewise lead to a cost in phonological processing, and that uniqueness point measures capture large changes in cumulative surprisal. As a consequence, a Bayesian approach in which only the point in time at which a word reaches a probability threshold determines response latencies must be incomplete, as in such an approach the cognitive cost of updating the probability distribution of lexical candidates is ignored.

Fig. 3 illustrates the time course of the posterior probabilities of the five sequences (left panel), and the corresponding time course of cumulative surprisal (right panel). Strings *abde* and *abdef* become unique at the same timestep, after 5 elements have become available in the input. For string *abde*, however, the increase in probability from position 4 to 5 is much larger than for string *abdef*. The right-hand panel, shows that the cumulative surprisal for *abdef* at position 5 is smaller than that of *abde*. If surprisal is indeed a measure of a cognitive processing bottleneck, then the processing cost of *abde* must be larger than that of *abdef*, even though both reach the maximal posterior probability simultaneously.

Cumulative surprisal functions depend on the similarity structure of the instance base (lexicon or grammar), as well as on the a priori probabilities of the elements (segments or words). In order to obtain more realistic cumulative surprisal functions, we calculated cumulative surprisal functions for 19,902 Dutch word forms (3584 monomorphemic, 3000 suffixed, and 13,318 compound words) of restricted lengths in a lexicon of 93,013 word forms. We turned to Dutch, rather than to Danish, because calculations critically depend on a phonological and morphological information in a database such as CELEX, which is not available for Danish. Since Dutch and Danish are related Germanic languages with similar morphologies, the results for Dutch provide a first approximation for Danish.

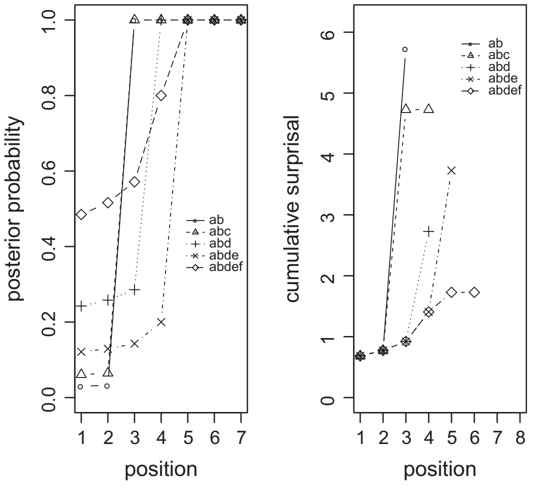

Fig. 3. Posterior probability (left) and cumulative surprisal (right) for the strings in the example universe of Table 5 beginning with *ab*.

Fig. 4 plots the average cumulative surprisal for Dutch monomorphemic words (top panels), suffixed words (central panels), and compounds (lower panels), for selected representative word lengths. The different curves in a panel represent the average trajectory of the cumulative surprisal for subsets of words with the uniqueness point occurring at the same position in the phoneme sequence. On each trajectory, the uniqueness point shared by the words represented on that trajectory is marked at the position where the uniqueness point is reached. The uniqueness point is calculated here in exactly the same way for each panel, following its original definition as the point at which the first constituent (or only constituent, in the case of simple words) becomes unique, disregarding morphological continuation forms (i.e. corresponding to our UP1). Surprisal, by contrast, is calculated relative to the full lexicon.

First consider the top two panels. For simple words, we see that once the uniqueness point has been reached, the cumulative surprisal asymptotes abruptly. This suggests that at the uniqueness point, the summed processing costs of a word have reached their near-maximum. The earlier this uniqueness point is reached, the earlier the processing costs required for disambiguation have been invested, and the earlier a lexical decision response can be initiated. This pattern is consistent with the effect of the uniqueness point on lexical decision latencies. Note, furthermore, that for longer words illustrated in the top right panel, the asymptotic cumulative surprisal increases relative to the shorter words shown in the top left panel. This is consistent with a positive correlation of word length and response latency.

The remaining panels illustrate that the classical uniqueness point no longer marks such a clear discontinuity in the cumulative surprisal function for complex words. After the uniqueness point of the first constituent has been reached, disambiguation is not yet completed, due to competition from morphologically related continuation forms. For the suffixed words (central panels), a slight increase after the uniqueness point is visible. Since many suffixed words are inflected words, and as the possibilities for building complex words on top of inflected words are severely restricted, the number of morphological continuation forms for suffixed words is limited. As a consequence, asymptotic levels of cumulative surprisal are close to the cumulative surprisal reached at the uniqueness point for suffixed words.

For compound words (lower panels), we observe a more pronounced increase in cumulative surprisal after the uniqueness point, notably for compounds with early uniqueness points, consistent with the effect of the CUP in the present study. Finally note that for most panels, the increase in cumulative surprisal (i.e., the surprisal of the segment at a given position) is greatest at the uniqueness point, even for complex words.

Given our interpretation of the UP effects as a result of large shifts in relative entropy, i.e. large changes in surprisal, we wanted to test whether some version of the surprisal measure would be a significant predictor of response latency in Experiment 1. In this connection, we are faced

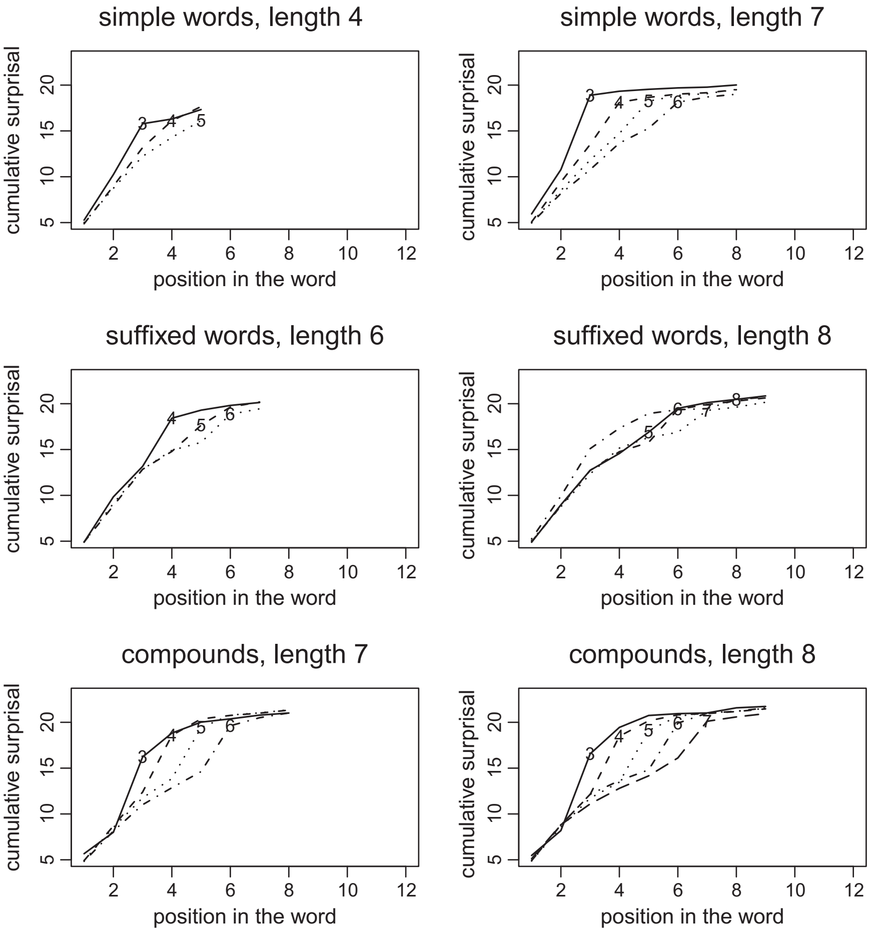

Fig. 4. Cumulative surprisal for monomorphemic words of lengths 4 and 7 (upper panels), suffixed words of lengths 6 and 8 (central panels), and compounds of lengths 7 and 8 (lower panels) in Dutch. Each curve represents the average for all words sharing the same uniqueness point, and is labeled with this uniqueness point.

with two problems: The first problem is that we have no sufficiently large phonologically transcribed corpus of Danish. Therefore, we were forced to use our written corpus to calculate surprisals, giving us a measure of surprisal for the orthographic form, which provides a rough index of the surprisal in the phonological forms. Our second problem is that raw surprisal values cannot be used as a predictor, since for each individual word there are as many surprisal values as there are segments. Therefore, we took as our starting point the cumulative surprisal—the measure depicted in Fig. 4—and fitted a linear regression to the cumulative surprisal function for each word, using the slope of this regression line as a predictor. If high surprisal comes early in the word, the slope of the regression line is steep, if surprisal comes later in the word, the slope is shallower.

This is a rather crude measure, but nonetheless our model for the response latencies in Experiment 1 is significantly improved by the addition of cumulative surprisal slope as an additional predictor. Considered by itself, as a main effect, the effect of the cumulative surprisal slope is non-linear, with the lowest reaction times observed for median values of cumulative surprisal slope, while both shallow (late high surprisal) and, especially, steep slopes (early high surprisal) gave rise to long response latencies. Since this suggests response optimization to the surprisal values that are most typical in the experiment, we investigated whether the cumulative surprisal slope differed as a function of any of the predictors that index participants' progress through the experiment. We observed a highly significant interaction between trial number and cumulative surprisal slope which is depicted in Fig. 5.[1] The tensor product modeling this interaction was supported by an analysis of deviance test ($F = 2.924$, $p = 0.0127$). (The effects of the other predictors in the model remained virtually unchanged, and therefore are not reported again.)

Fig. 5 illustrates the relation between trial number (on the horizontal axis) and the cumulative surprisal slope (on the vertical axis). Longer response latencies are indi-

[1] Our main goal here is to test the Shortlist B model of auditory word recognition, but an alternative model also presented by Norris and McQueen (2008) is the Merge B model in which evidence for all phoneme positions is allowed to accumulate throughout the word. If initial changes in relative entropy are weighted more strongly than later surprisals, as we surmise would be the prediction of the Merge B model due to the continued accrual of evidence over time, this does not change the pattern we observe for the cumulative surprisal slopes.

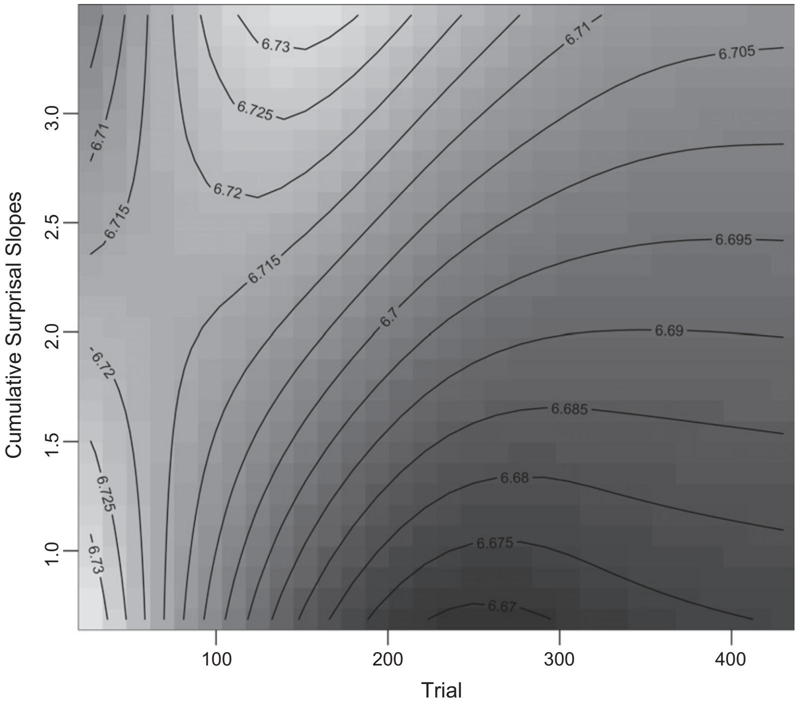

Fig. 5. Partial effect of Trial by cumulative surprisal slopes using a tensor product. Lighter shades of gray indicate longer log response latencies. Contour lines show log RT values; changes in RT are larger when contour lines are close.

cated by lighter shades of gray. What this plot shows is that early in the experiment, participants are slowest on the shallow surprisal slopes, i.e. those words in which the highest surprisal comes late. This pattern quite quickly begins to shift such that words with steeper surprisal slopes, indicating early high surprisal, become the most difficult, peaking around trial 150. Later in the experiment, there is less effect of surprisal but words with the steepest slopes remain the most difficult.

This pattern suggests that participants are sensitive to how surprisal is distributed in the words they hear, in relation to the other words heard in the experiment, i.e. participants are optimizing their responses in the context of the experiment. Participants enter the experiment not expecting shallow surprisal slopes, and then rapidly change their expectations (there are many complex words in the experiment, which tend to have relatively shallow surprisal slopes), and proceed to optimize their performance as the experiment proceeds. Even though the cumulative surprisal slopes are relative crude measures, they show that participants' information processing in auditory word recognition is sensitive to the surprisal.

Interestingly, the UP-measure remains stable and highly significant also when the cumulative surprisal slopes are included in the model. We see at least two reasons for this: Firstly and most trivially, the UP-measures are relatively fine descriptions of the acoustic signal that participants hear, whereas the surprisal slopes are based on the orthography and are simplified linear descriptions of the cumulative surprisal curves such as those illustrated in Fig. 4. This is also at least part of the reason why the effect size for the UPs are on the order of 300 ms (see Fig. 2), while the maximal effect size for cumulative surprisal is 48 ms (the difference between the lowest value in Fig. 5, 6.67 log RT, and the highest, 6.73 log RT). More importantly, the presence of both effects reflects a fundamental difference between the two measures: although related, the surprisal slopes probe how cumulative information is processed, while the UP-measures index the important change points in cumulative surprisal, points that reflect the distributional properties of the language.

Experiment 1 documented inhibitory effects of UP1 and CUP, no additional morphological family size effect, and a facilitatory effect of the number of continuation forms. Experiment 2 presents a replication study with new materials, replacing the compound words by suffixed words. As the distributional survey of Dutch suggests that compounds might show the clearest increments in cumulative surprisal after UP1, removal of compounds as stimuli makes it more difficult to detect an effect of CUP. Experiment 2 will also allow us to ascertain whether the facilitatory effect of the continuations count is robust. Finally, we expect to observe an effect of cumulative surprisal slope, and we expect it to again interact with Trial.

4. Experiment 2

4.1. Method

4.1.1. Materials

We selected 125 Danish derived forms for presentation in auditory lexical decision. Additionally, the experiments

Table 6
Lexical predictors for the items in Experiment 2 and 3. The variables marked with an asterisk are frequency counts per million words based on a 43.6 million word corpus of Danish. All other variables are based on the same corpus. $N = 125$.

Predictor	Mean	SD	Range
UP1, ms	289	85	131–579
Complex UP, ms	456	103	232–732
Duration, ms	645	97	458–940
Length in phonemes	7.0	1.3	4–10
Length in letters	7.8	1.5	4–12
Continuation forms (type frequency)	12	19	1–114
Cohort density	288,630	292,555	2801–1,338,611
Neighborhood size	0.5	0.9	0–4
Whole-word frequency*	6	12	0–100
Stem frequency*	398	2904	0.2–32,437
Family size	341	405	1–2103
Affix type frequency	11	13	0.4–51
Mean bigram frequency*	30,951	16,076	4581–88,334
Juncture bigram frequency*	1552	21,625	72–142,493

included 110 simple and 15 compound word fillers, and 250 nonwords. The items were selected in a similar way to those of Experiment 1. The derived words were a mixture of suffixed words, prefixed words, and words carrying particle prefixes. There were ten suffixes, seven prefixes, and eight particle prefixes, each of which occurred on five words in the experiment. Some of the affixes were also included in Experiment 1 while only three items were repeated from Experiment 1; all items are listed in the appendix. All affixes were relatively productive and none were homonymic with other affixes (Bertram, Laine, Baayen, Schreuder, & Hyönä, 1999; Bertram, Laine, & Karvinen, 1999), though the verbalising suffix *-ere* is homographic (but not homophonic) with the comparative. All words carrying this affix were unambiguously derived verbs. Predictors for the items are summarized in Table 6. The 250 nonwords were constructed by changing one to three phonemes in each word, while retaining the affixes of the complex words on the nonwords.

The stimuli were read by the same female native speaker of Danish as those for Experiment 1. The stimuli were recorded on a Sony DAT-recorder (model TCD-D8), using a Sony electret condenser microphone (model EC-959a), in a sound-attenuated room. The recordings were digitized at a sampling rate of 22 kHz and a bit depth of 16 bit. Items were normalized for peak intensity.

4.1.2. Participants

21 volunteers (11 women and 10 men between the ages of 22 and 39, mean 26.3 years) participated in the experiment. The participants were from the same population as those in Experiment 1.

4.1.3. Procedure

The procedure was identical to the one used in Experiment 1, except that ISI was fixed at 3000 ms for all participants. The experiment lasted about 30 min.

4.2. Results and discussion

Errors constituted 4.3% of the lexical decision responses; these were excluded from the RT analyses. Additionally, due to error rates over 30%, the responses to four items were removed for the RT-analyses. All in all, 5.4% of responses were excluded from the RT-analysis due to errors, while 2.7% of the remaining datapoints were excluded due to large standardized residuals as in Experiment 1. All data points were retained for the error analysis. Variables were logarithmically transformed as in Experiment 1. The data were analyzed using generalized additive mixed models in the same manner as for Experiment 1; the results are shown in Tables 7–10.

As for Experiment 1, decorrelation of some variables was necessary: The RTs on the four previous trials were orthogonalized using Principal Components Analysis. The overlapping durational measures UP1, CUP, and word duration were recalculated as non-overlapping parts of the signal. The number of continuation forms was residualized from whole-word frequency and stem frequency was residualized from morphological family size. These decorrelated variables are marked as ‘‘Residualized’’ or ‘‘Resid’’ in the tables and figures. With these decorrelated measures, collinearity was low, with a condition number κ below 10 for both models.

4.2.1. Control predictors

As for Experiment 1, the principal components introduced to remove distortions due to autocorrelational structure in the time series of responses reached significance, with the usual large effect sizes.

4.3. Uniqueness points and surprisal

This experiment shows effects of both UP1 and duration from UP1 to CUP. Both effects were linear. There was no effect of the duration from CUP to offset, the measure that is our index of word duration. However, both UPs remained significant irrespective of whether the decorrelated or the full duration measure were included in the analysis. This experiment provides clear confirmation of Experiment 1 and of Balling and Baayen (2008): we see strong and significant effects of both UP1 and CUP for all types of complex words, with no significant differences between different types of complex words.

Table 7
Parametric coefficients of the generalized additive mixed model fitted to the response latencies of Experiment 2.

	Estimate	Standard Error	*t*	*p*
Intercept	5.4668	0.1900	28.7782	0.0000
Previous RT PC1	0.1114	0.0119	9.4024	0.0000
Previous RT PC2	0.0627	0.0132	4.7448	0.0000
Log UP1	0.1778	0.0292	6.0953	0.0000
Log UP1 to CUP	0.0324	0.0052	6.1889	0.0000
Log Cohort Sum Frequency	0.0161	0.0063	2.5572	0.0106
Log Family Size	0.0179	0.0049	3.6260	0.0003

Table 8
Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), *F* and *p* values for the tensor products and random effects in the generalized additive mixed model fitted to the response latencies of Experiment 2.

	edf	Ref.df	*F*	*p*
tensor Log Word Freq by Resid Stem Freq	13.6686	16.5292	3.2870	0.0000
tensor Cumulative Surprisal Slope by Trial	3.0172	3.0204	24.8849	0.0000
random intercepts Word	19.6721	19.9937	30.4665	0.0000
random intercepts Participant	89.8758	106.4264	3.3061	0.0000

Table 9
Parametric coefficients of the generalized additive mixed model fitted to the response correctness for Experiment 2.

	Estimate	Standard error	*z*	*p*
Intercept	−19.1713	4.3595	−4.3976	0.0000
Log CUP	2.6168	0.5870	4.4581	0.0000
Log Word Frequency	−0.3428	0.0707	−4.8503	0.0000
Log Continuations	−0.9161	0.1474	−6.2170	0.0000
Affix:Prefix	2.4159	0.3144	7.6852	0.0000
Affix:Particle	1.7725	0.3140	5.6441	0.0000

Table 10
Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), *F* and *p* values for the random effect of Participant in the generalized additive mixed model fitted to the response correctness for Experiment 2.

	edf	Ref.df	Chi.sq	*p*
random intercepts Participant	19.4447	19.9104	38.0208	0.0085

The effect of UP1 was large and linear, as can be seen in Fig. 6. Except for a small number of low values, the effect of log UP1 to CUP was linear as well, with a somewhat smaller effect size. The linear effect of UP1 in this experiment replicates the linear effect of UP1 observed in Balling and Baayen (2008) for an auditory lexical decision experiment with only suffixed words. The non-linear effects in Experiment 1 may be due to the presence of compounds in the experimental lists, which is also the main reason for the longer item durations in Experiment 1 than in Experiment 2. Possibly, the cognitive costs of surprisal increase with each successive segment that becomes available in the input, due to increasing demands on working memory for increasing numbers of segments and morphemes. Due to the reduced length of the items in Experiment 2, the UP effects would then be linear.

A cohort density measure, the log of the summed frequencies of the members of the onset cohort, also reached significance, unlike in Experiment 1. The original measure of Magnuson et al. (2007), the sum of the log-transformed frequencies, did not reach significance. The effect of cohort density was non-linear: the fastest response latencies were observed for words with intermediate cohort densities, while words with high and low densities were responded to more slowly. This U-shaped pattern (see Fig. 6) suggests response optimization for the more likely, central-valued cohort densities, coming with the price of longer responses for more peripheral, extreme, low-probability cohort densities (cf. the effect of cumulative surprisal slope and Tabak, Schreuder, & Baayen, 2010). Importantly, both UPs remain significant with cohort density as a co-predictor in the statistical model. No effect of neighborhood density could be observed.

In addition to the UP-effects, we again observe an interaction between cumulative surprisal slope and trial, which is depicted in the left panel of Fig. 7. The overall shape of this interaction suggests that words with median-valued surprisal slopes, around 2, are the fastest to process—here we observe the darkest shades of gray. Throughout the experiment, the steep surprisal slopes (corresponding to

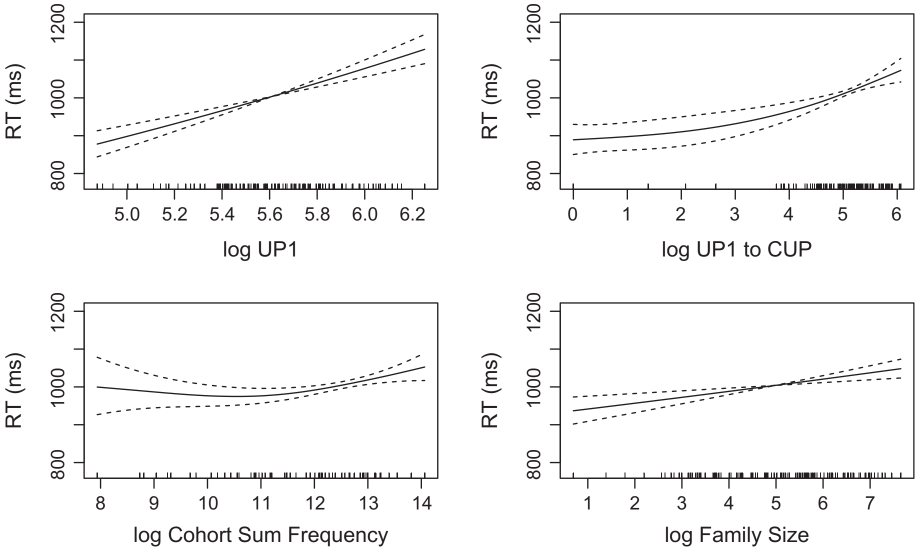

Fig. 6. Partial effects in Experiment 2 of log UP1, log UP1 to CUP, log cohort sum frequency, and log family size (adjusted for the intercept, and back-transformed from the log scale to ms scale). Due to identifiability constraints, linear predictors have zero confidence intervals where the (unshifted) partial effect is zero.

early high surprisal) are the most difficult, indicated by lighter shades of gray. Furthermore, words with shallow surprisal slopes also tended to elicit longer response latencies. Overall, as subjects proceeded through the experiment, response latencies tended to increase, except for words with very steep cumulative surprisal slopes, for which latencies tended to decrease as the experiment progressed. This suggests participants are again engaging in a process of response optimization, allowing them to deal more effectively with words with steep surprisal slopes, at the cost of words with shallow surprisal slopes. The shape of this response optimization is different from Experiment 1, probably because the shorter words in Experiment 2 rendered the experiment less demanding overall.

4.4. Frequency

The RT-analysis showed an interaction between stem and whole-word frequency, depicted in the right panel of Fig. 7. Whole-word frequency has a facilitatory effect throughout the experiment, but it is strongest for words with a low residualized stem frequency and attenuated for those with higher stem frequency. Stem frequency only has an effect for words with lower whole-word frequency.

The shape of this interaction between whole-word and stem frequency is remarkably similar to the interaction between whole-word and affix frequency observed by Balling and Baayen (2008). For the suffixed words of Balling and Baayen (2008), this corresponds to an interaction between whole-word and second constituent frequency. In the pres-

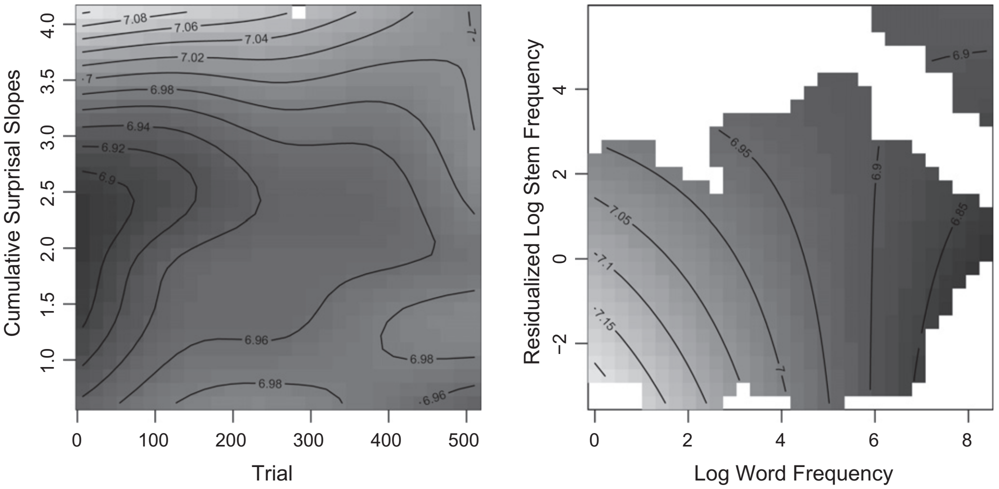

Fig. 7. Partial effect of Trial by cumulative surprisal slopes and log word frequency by residualized stem frequency using tensor products. Lighter shades of gray indicate longer log response latencies.

ent experiments, the majority of the complex words were prefixed, suggesting that the interaction also in this case is between whole-word and second constituent frequency. This is confirmed when suffixed and prefixed words are analyzed separately: The interaction between stem and whole-word frequency remained significant for the prefixed words, but was completely absent for the suffixed words when analyzed alone. Although a three-way interaction between affix type, whole-word and stem frequency did not reach significance, it seems safe to conclude that the frequency interaction is primarily carried by the prefixed words. This means that the present experiment and the auditory lexical decision experiment of Balling and Baayen (2008) are consistent in showing an interaction between whole-word and second constituent frequency. Similar interactions between whole-word and constituent frequencies are reported by Kuperman et al. (2009) and Kuperman, Bertram, and Baayen (2010).

For auditory comprehension, this interaction may fall out from the dynamics of lexical competition in models such as Shortlist B, with a trade-off between bottom-up support and a priori probability leading to shorter words (with less bottom-up support) with higher frequencies dominating longer words (with more bottom-up support) with lower frequencies. Without a computational implementation of Shortlist B for Danish, we cannot further test this hypothesis, unfortunately.

4.5. Family size

The family size count measure was significant, and inhibitory, in Experiment 2 (see Fig. 6). For reasons unclear to us, Experiment 2 was more successful than Experiment 1 in picking up competition effects, not only from the morphological family, but also from the (modified) cohort density of Magnuson et al. (2007). In contrast to Experiment 1, the count of continuation forms did not reach significance.

Why is the family size effect inhibitory in this experiment, but facilitatory in visual lexical decision? As discussed above, one important factor is likely to be the fact that auditory processing proceeds sequentially from word onset to offset, which means that onset-alignment—which is disregarded in the classic family size count—is likely to influence word recognition more in the auditory than in the visual modality. In order to explore this further, we conducted a supplementary analysis of the suffixed words, for which we could straightforwardly divide the already manually sorted families into those members that are not onset-aligned with the target, those that are continuations of the first constituent (i.e. those that are potentially part of the competition indexed by the CUP), and those that are continuations of the target whole word.

The first thing we observed is a high correlation between the full family size count and the count of non-onset-aligned family members ($R = 0.95$). This suggests that the inhibitory family effect may well be driven by the non-aligned members. When including the two counts of aligned and non-aligned family members in the regression model, we see that the non-aligned part of the family carries the inhibitory effect.

If this inhibitory effect of non-aligned family members turns out to be replicable, the way in which Shortlist B currently allows non-onset aligned competitors into its calculations is too restrictive. For a target word such as *logbook*, *book* is taken into account as competitor for the input string *logb*, but *handbook* is not. Experiment 2 suggests that once sufficient evidence for *book* has accumulated, *handbook* has also become a competitor with a measurable negative effect on response speed.

The remaining part of the full family, those words that are onset-aligned with the target and overlap in one morpheme show no significant effect, but these words are already included in the analysis in the sense that these are the words from which the target becomes unique at the CUP. As expected, the temporal CUP-measure (UP1 to CUP duration) remains significant in this subanalysis.

Another way of partitioning the set of family members is by the frequency of the family members compared to the frequency of the target. Meunier and Segui (1999) reported that words were harder to recognize when they had many family members of higher frequency than the target frequency. A division between higher and lower frequency family members was not informative for our data; neither variable was significant on its own. We conclude that, at least for the present Danish data, this partitioning of the family is not helpful.

We have argued that inhibition for the morphological family size measure in the auditory modality is due to most family members being non-onset-aligned. By implication, and given the facilitatory family size effects reported for visual comprehension, the stimuli of Experiment 2 should elicit a facilitatory family size effect in visual lexical decision. This prediction is tested in Experiment 3.

Additionally, Experiment 3 investigates the role of UPs and cumulative surprisal slopes. There are two reasons for testing such inherently auditory measures in a visual experiment: Firstly, these measures are central to our understanding of auditory word recognition and it is therefore important whether one or more of them are also observed in the visual modality (as is the case for the standard UP in the visual lexical decision task of Baayen et al., 2007). Secondly, any role for such auditory measures in visual lexical decision would be very informative about the relation between the two modalities and an important fact for models of reading to account for.

5. Experiment 3

5.1. Method

5.1.1. Materials

The same materials were used as in Experiment 2.

5.1.2. Participants

20 participants (12 women and 8 men between the ages of 21 and 38, mean 26.8 years) from the same population as in the previous experiments participated in Experiment 3. All reported normal or corrected-to-normal vision.

5.1.3. Procedure

The experiment was run on the same equipment as the two previous auditory experiments, but stimuli were presented visually. Each stimulus was preceded by a plus in the middle of the screen which was replaced with the stimulus centered around the same point after 500 ms. The items were white on a black background, presented in a lower-case Courier New 18 point font on a 15 inch screen. Each item was displayed for 2000 ms or until the participant responded, giving an ISI which was variable but with a maximum of 2500 ms. The experiment lasted 10–15 min.

5.2. Results and discussion

Error responses (5.9%) were excluded from the analysis of RT, along with all responses to five items with error rates above 30%. In this way, 7.0% of responses were excluded from the RT analysis, while all responses were retained for the error analysis. As for the previous experiments, analyses were conducted using generalized additive mixed-effects regression models. The RT analysis is summarized in Tables 11 and 12 and the error analysis in Tables 13 and 14. The RT model was again trimmed to exclude overly influential outliers, removing 2.5% of the datapoints. The steps taken to reduce collinearity for Experiment 3 are also relevant here; additionally, word length in letters was residualized from whole-word frequency and juncture bigram frequency from stem frequency. With these decorrelated variables, κ was below 10.

5.2.1. Control variables

In this experiment, only the first of the principal components based on the RTs on previous trials had a significant effect on current RT. Latencies were longer, and errors less probable, when preceding trials had longer latencies. Moreover, RTs were significantly longer when the preceding response had been incorrect than when it had been correct.

5.2.2. Uniqueness points

The analysis showed a significant effect of UP1 in ms which entered into an interaction with cumulative surprisal slope. Conceptually, both are descriptions of the temporal unfolding of the auditory input. It is therefore somewhat surprising that we see effects at all, though as noted above, Baayen et al. (2007) found effects of the standard UP in a visual task. Importantly, this is not an artefact of a letter-based UP1: there were no effects of visual UP1 or CUP, neither linear or non-linear effects. Furthermore, the temporal measures based on the auditory stimulus remained significant, even when the visual measures were added to the model. In the error analysis, UP1 also showed an inhibitory effect: words with late UP1 were more error-prone, confirming the result of the RT-analysis.

As can be seen in Fig. 8, there is a main effect of UP1: response latencies increase with increasing UP1. Interestingly, the inhibitory effect of the UP1 is most pronounced for median (a priori most likely) values of the cumulative surprisal slope. Furthermore, for values of UP1 below 5.7, the contour plot suggests that words with unexpectedly high or low values of the cumulative surprisal slope elicit longer response latencies. This suggests that in reading, response optimization is achieved by conditioning on the expectation that a word will have an average (unsurprising) cumulative surprisal slope.

The effects of UPs and cumulative surprisal are compatible with the idea of auditory recoding of the visual input in reading (e.g. Lukatela, Eaton, Lee, Carello, & Turvey, 2002). If the UP effect does indeed indicate a mental replaying of the word presented visually, the interaction suggests that such replaying is most succesful for the items with the most typical surprisal slopes. The significance of an auditory UP1 measure as predictor for visual comprehension observed in the present study, replicating the result obtained by Baayen et al. (2007) for English, challenges the counterpart of Shortlist B for reading, the Bayesian Reader (Norris, 2006), as well as connectionist models such as the DRC model of Coltheart, Rastle, Perry, Langdon, and Ziegler (2001).

5.2.3. Word length

Word length in letters showed the expected inhibitory effect with longer response time to longer words.

5.2.4. Frequency

There was a significant effect of whole-word frequency as well as a reduced but significant effect of stem frequency. The word frequency effect was about twice the size of the stem frequency effect, replicating other studies reporting such joint frequency effects for reading (e.g., Baayen et al., 1997, 2007; Kuperman et al., 2010).

Additionally, visual decision latencies were affected by the frequency of the letter bigram crossing the morpheme boundary: the less frequent the bigram and consequently the more noticeable the morpheme boundary, the faster

Table 11
Parametric coefficients of the generalized additive mixed model fitted to the response latencies of Experiment 3.

	Estimate	Standard error	t	p
Intercept	6.6777	0.0234	285.6938	0.0000
Previous RT PC1	−0.1198	0.0133	−8.9821	0.0000
Previous Response:Error	0.0897	0.0191	4.7023	0.0000
Residualized Length in Letters	0.0226	0.0061	3.6998	0.0002
Log Word Frequency	−0.0353	0.0040	−8.7442	0.0000
Residualized Stem Frequency	−0.0138	0.0029	−4.7388	0.0000
Residualized Juncture Bigram Freq	0.0105	0.0027	3.8883	0.0001
Residualized Continuations	−0.0297	0.0053	−5.6080	0.0000

Table 12
Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), *F* and *p* values for the splines, tensor products and random intercepts for Participant (random intercepts for Item were not supported) in the generalized additive mixed model fitted to the response latencies of Experiment 3.

	edf	Ref.df	*F*	*p*
splines Log Family Size	3.0480	3.8367	5.7283	0.0002
tensor product Cumulative Surprisal by UP1	8.6763	9.9090	3.8145	0.0000
random intercepts Participant	17.6591	18.9043	7.8606	0.0000

Table 13
Parametric coefficients of the generalized additive mixed model fitted to the response correctness for Experiment 3.

	Estimate	Standard error	*z*	*p*
Intercept	−5.4234	1.7571	−3.0865	0.0020
Previous RT PC1	0.9170	0.2765	3.3159	0.0009
Log UP1 ms	0.9752	0.2996	3.2551	0.0011
Log Word Frequency	−0.2971	0.0566	−5.2492	0.0000
Log Continuations	−0.6557	0.1180	−5.5564	0.0000
Log Family Size	−0.2113	0.0529	−3.9948	0.0001

Table 14
Estimated degrees of freedom (edf), reference degrees of freedom (Ref.df), *F* and *p* values for the random effect of Participant in the generalized additive mixed model fitted to the response correctness for Experiment 3.

	edf	Ref.df	Chi.sq	*p*
random intercepts Participant	13.4284	17.2718	32.1344	0.0161

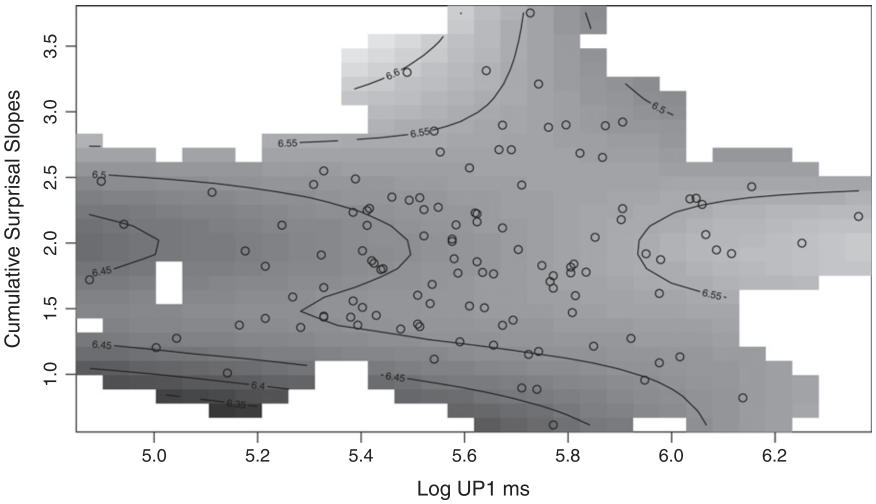

Fig. 8. Partial effect of UP1 by cumulative surprisal slopes using a tensor product. Lighter shades indicate longer log response latencies; circles indicate items, positioned according to their values on the two predictors.

the reaction time. This suggests that morphological information is more available in reading when a boundary is more clearly marked, facilitating word recognition.

5.2.5. Family size

The left panel of Fig. 9 visualizes the non-linear effect of morphological family size, which is similar to that observed by Baayen, Feldman, and Schreuder (2006) and Tabak, Schreuder, and Baayen (2005): Up to a certain point, more family members help recognition, while above that point, the effect becomes inhibitory. It seems that when a very large number of family members are activated, they constitute noise that begins to inhibit the recognition of the target. Moreover, large families also tend to be semantically more diverse, which should detract from the facilitatory family effect if it is indeed a semantic effect as argued by, among others, De Jong et al. (2000). This is confirmed by the fact that both in the auditory Experiment 2

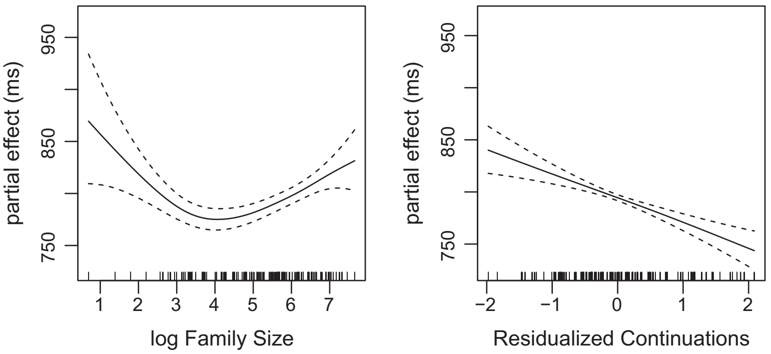

Fig. 9. Partial effects of log family size (using restricted cubic splines) and residualized continuations, shifted by the intercept and back-transformed from the log scale to the ms scale.

and in the visual Experiment 3, the semantically least related family members were responsible for most of the inhibition. Here, we used what Moscoso del Prado Martín et al. (2004) term the non-dominated family—those words that are family members of the stem rather than the whole complex word—to index the semantically least related family members. This inhibition is also reminiscent of the secondary family size effect reported in Baayen (2010): the family members of family members appear to have a small inhibitory effect in visual lexical decision. These secondary family members are also words that are semantically distant (e.g., *tea trolley* and *bus stop* linked through *trolley bus*). In contrast to the traditional family size, continuation forms had a straightforwardly linear facilitatory effect, as in Experiment 1, which is illustrated in the right panel of Fig. 9.

The fact that morphological families in Danish are substantially larger than those in Dutch and English may also contribute to the strong support for significant non-linearity. In the error analysis, the inhibitory component is absent and we observe that words with larger families are less error prone.

6. General discussion

Experiments 1 and 2, as well as the experiment reported by Balling and Baayen (2008), provide strong evidence that there are two points during the auditory recognition of a word at which its probability shifts dramatically: first, when unrelated words become incompatible with the target at UP1; next, when words that share the first morpheme are no longer possible candidates at the CUP. Together, the two UP-measures provide a way of gauging the development of lexical probabilities over time in a complex dynamic system like the one underlying auditory comprehension, based on lexical statistics rather than computational implementations.

We have argued that the presence of an effect of UP1 for complex words is unexpected within models such as Shortlist B. Instead, in Shortlist B, a lexical decision is proportional to the point in time at which the evidence in the input is fully covered by lexical representations with posterior probabilities close to 1; this point in time will be in the vicinity of the CUP. We think this point in time is indeed important, as documented by the significant effects of the CUP across three auditory experiments. However, we see two problems with relying on the CUP alone as a predictor of reduction of the cohort: Firstly, the UP1, which indexes competition from non-related cohort members, has at least as strong effects as the CUP, showing that there are two points in the processing of bimorphemic words at which the cohort probabilities shift substantially. Secondly, the CUP as it would be interpreted within Shortlist B lacks precision as a measure of processing cost because it is assumed (implicitly) that the process of updating probabilities with each additional phoneme that becomes available to the listener is cost-free. It is conceivable that a large revision should be cognitively more costly than a small revision. Interestingly, such a cost is found not only for updating during sentential reading, by Levy (2008) and others, but in this study also for updating in auditory comprehension.

The fact that both UP1 and CUP are significant predictors with substantial effect sizes indicates that the cognitive costs of updating from a probability distribution d_t to a new probability distribution d_{t+1} can be substantial and should be brought into Bayesian models of auditory comprehension.

Also supporting the role of surprisal is the evidence that the recognition of a word is influenced by how cumulative surprisal develops within the word. Though both UPs and cumulative surprisal slopes document the importance of large changes of probability within the competition cohort, there are also differences: The cumulative surprisal slopes provide a general measure of how surprisal develops across the word. The UPs, by contrast, focus on time, measuring two points in time at which the cohort changes substantially. These are, of course, measures of formal overlap, but by also taking morphological structure into account we now index in addition the points where the semantic structure of the competition space changes: The members of the cohort become much more semantically uniform, especially at the UP1 where all morphologically unrelated words disappear from the cohort, but also at the CUP after which only continuation forms that fully contain the target remain compatible with the input.

Our cumulative surprisal slopes are estimated for unambiguous input. If one or more segments is ambiguous, we expect surprisal to shift to a later stage in the word where the listener has become certain enough about the input to reduce the competition cohort. Hence, the present cumulative surprisal slopes provide upper boundary estimates.

A surprising finding in this study is that UP1 is predictive not only for auditory lexical decision, but also for visual lexical decision, as demonstrated by Experiment 3. This result, which is compatible with the ontogenetic primacy of the auditory modality and with previous evidence of the influence of auditory information on visual processing (Lukatela et al., 2002; Baayen et al., 2006; Baayen et al., 2007), challenges all current models of reading, whether Bayesian (Norris, 2006) or connectionist (Coltheart et al., 2001).

We expect that further explanatory power might be gained by bringing subphonemic cues into the analyses, in addition to the phonemic UP-measures (Davis, Marslen-Wilson, & Gaskell, 2002; Salverda et al., 2003; Kemps, Ernestus, Schreuder, & Baayen, 2005). Doing so is beyond the scope of this paper, as it requires detailed information on subphonemic similarities and differences between the target words and all their competitors. Based on a comparison of the size of the UP-effects in the present experiments (200–350 ms, comparable to the effect size of word frequency, 300 ms for a frequency range of 0 to 100 per million) with those of the subphonemic effects reported by Kemps et al. (2005) (on the order of 60 ms) we would expect the precision gain to be minor.

Further refinements are also expected when competitors that are not onset-aligned with the target itself but are onset-aligned with a non-initial substring of the target, such as *rusty* for *crust*, are brought into the calculations of the probability distributions d_t and d_{t+1} and their relative entropy. The probability distributions generated by Shortlist B, which takes non-onset-aligned competitors into account, are therefore a better starting point for the calculation of cumulative surprisal than the onset-aligned calculations that have informed our research. Without a computational implementation of Shortlist B for Danish, we have not been able to pursue this line of research further. Given the large effect sizes of UP1 and CUP in our experiments, we expect more precise, but not substantially different effects to be obtained.

Our experiments also shed light on the role of the morphological family in auditory comprehension. The main pattern that emerges across Experiments 1 and 2 is that morphological family size generally has an inhibitory effect in auditory comprehension, as opposed to reading where facilitatory effects are observed. This inhibition emerges in two ways: firstly, the standard morphological family size effect is inhibitory in Experiment 2, though it does not reach significance in Experiment 1. Secondly, the CUP is also based on a subset of the morphological family, namely those words that share the target's first constituent as a first constituent, and this has a reliably inhibitory effect across both the auditory experiments presented here, as well as in the experiment reported by Balling and Baayen (2008). The only measure of morphological family size that has a facilitatory effect is the count of continuation forms, i.e. words that are both onset-aligned and fully overlap with the target, which emerges as a significant facilitatory predictor in Experiments 1 and 3. This suggests that their probability mass is pooled with that of the target word, even though at target offset the current implementation of Shortlist B would eliminate such competitors as incompatible with the evidence. Thus, it seems that onset-alignment affects which morphologically related word emerge as competitors and which result in facilitation. This is compatible with the analysis of suffixed words in Balling and Baayen (2008) where the fact that the shared constituent of the family begins at the onset of the suffixed target words probably contributed to the emergence of this effect as facilitatory. The importance of onset-alignment to competition dynamics in auditory comprehension may also contribute to the difference in these effects between auditory and visual comprehension, but further research is needed to fully uncover the role of morphological family members across modalities.

The UP1 and CUP measures are morphological measures in the sense that they assess critical points in the lexical disambiguation process in terms of morphological relatedness of competitors: morphologically totally unrelated at UP1, and morphologically partially related at CUP. The two measures were defined explicitly to generalize across prefixed words, compounds, and suffixed words. The absence of any interactions of the uniqueness point measures with morphological type indicates that these measures are successful as general measures of cognitive bottlenecks in auditory comprehension.

The significance of the CUP as a predictor is especially telling in the case of suffixed words and compounds. For these, in theory, participants could have made a lexical decision immediately at the point that the first constituent became unique, i.e., at the UP1, because none of the nonwords started with real stems. Nevertheless, the effect of the CUP is as strong for suffixed words and compounds as it is for prefixed words. This indicates that the UP-effects are recognition- rather than decision-related. Furthermore, the fact that the ISI manipulation in Experiment 1 does not affect any of the similarity variables suggests that they are not narrowly task-dependent.

In this study, we have argued for a rehabilitation of the uniqueness point concept originating with the work of Marslen-Wilson and colleagues in the framework of cohort theory (Marslen-Wilson, 1984; Marslen-Wilson & Welsh, 1978). Cohort theory has been found to be too restrictive, for instance, in its insistence on onset alignment. With the abandonment of cohort theory, the uniqueness point as a predictor for auditory comprehension became unfashionable as well. By way of example, there is not a single mention of uniqueness points in the presentation of the Shortlist B model (Norris & McQueen, 2008). Yet, as researchers in auditory comprehension well know, a word's uniqueness point is often one of the strongest predictors in auditory lexical decision tasks. We have shown that morphologically-sensitive redefined uniqueness points measures, calculated on the basis of a full-form lexicon, remain valid as highly significant predictors of substantial lexical processing costs. Furthermore, we have shown that uniqueness points can be understood theoretically as points in time at which a large cognitive effort, gauged by a segment's surprisal, is required to update probability distributions of lexical competitors. In addition to these single points in time at which probabilities shift dramatically, listeners are sensitive to the general development of surprisal across a word. Surprisal in sentence processing (Hale, 2001; Levy, 2008) is thus mirrored at the lexical level in auditory comprehension, suggesting fractal self-similarity of cognitive processing costs in sequence processing at very different levels of linguistic structure.

Acknowledgments

We are grateful to Ocke-Schwen Bohn and Lee Wurm for discussion and to Dennis Norris and one anonymous reviewer for comments on an earlier version of this article. This work was partially supported by an Alexander von Humboldt professorship awarded to the second author by the Alexander von Humboldt foundation.

Appendix A

Tables A1 and A2.

Table A1

Items from Experiment 1, with information about morphological type and part of speech as well as glosses for first and second constituents and whole-word translations. For affixes, the general meaning of the affix is given once. For the prefix *be-*, trans/dir refers to its function of making its stem transitive or adding directional meaning.

Word	Type	PoS	Gloss 1	Gloss 2	Translation
adresse-bog	Compound	N	address	book	address book
blod-rød	Compound	A	blood	red	blood red
blød-gøre	Compound	V	soft	make	to soften
bort-føre	Compound	V	away	take	to abduct
brev-veksle	Compound	V	letter	exchange	to exchange letters
damp-koge	Compound	V	steam	boil	to steam
fast-spænde	Compound	V	firm	fasten	to fasten
fod-bold	Compound	N	foot	ball	football
frost-kold	Compound	A	frost	cold	ice cold
gade-dreng	Compound	N	street	boy	street urchin
gen-splejse	Compound	V	gene	splice	to genetically engineer
guld-medalje	Compound	N	gold	medal	gold medal
hals-hugge	Compound	V	neck	cut	to decapitate
hjerne-vaske	Compound	V	brain	wash	to brainwash
hyle-tone	Compound	N	wail	tone	high-pitched noise
kontor-chef	Compound	N	office	chief	head of department
kunst-historie	Compound	N	art	history	art history
lager-plads	Compound	N	storage	space	storage space
lokal-radio	Compound	N	local	radio	local radio
luft-tørre	Compound	V	air	dry	to air-dry
luksus-vare	Compound	N	luxury	item	luxury item
lys-stråle	Compound	N	light	beam	ray of light
mange-doble	Compound	V	many	double	to multiply
menneske-masse	Compound	N	human	mass	crowd
møbel-fabrik	Compound	N	furniture	factory	furniture factory
mål-rette	Compound	V	aim	direct	to aim
navn-give	Compound	V	name	give	to name
pande-hår	Compound	N	forehead	hair	fringe
pant-sætte	Compound	V	pawn	place	to pawn
plan-lægge	Compound	V	plan	lay	to plan
prøve-smage	Compound	V	trial	taste	to sample (something edible)
rask-melde	Compound	V	healthy	inform	to report fit for duty
risiko-gruppe	Compound	N	risk	group	risk group
selv-sikker	Compound	A	self	secure	self confident
silke-glat	Compound	A	silk	smooth	smooth as silk
slæde-hund	Compound	N	sledge	dog	sledge dog
små-snakke	Compound	V	small	talk	to chat
sne-blind	Compound	A	snow	blind	snowblind
sol-bade	Compound	V	sun	bathe	to sunbathe
spare-penge	Compound	N	save	money	savings
strand-vejr	Compound	N	beach	weather	beach weather
struktur-problem	Compound	N	structure	problem	structural problem
studie-lån	Compound	N	study	loan	student loan
styrke-træne	Compound	V	strength	practice	to work out
succes-rig	Compound	A	success	rich	successful
vakuum-pakke	Compound	V	vacuum	pack	to vacuumpack

(*continued on next page*)

Table A1 (*continued*)

Word	Type	PoS	Gloss 1	Gloss 2	Translation
videre-sælge	Compound	V	onwards	sell	to resell
vild-lede	Compound	V	wild	lead	to mislead
vind-mølle	Compound	N	wind	mill	windmill
æble-most	Compound	N	apple	juice	applejuice
af-blege	Particle	V	off/away	bleach	to bleach
af-dæmpe	Particle	V	–	curb	to subdue
af-grænse	Particle	V	–	limit	to limit
af-høre	Particle	V	–	hear	to interrogate
af-klare	Particle	V	–	clear	to clarify
af-kræve	Particle	V	–	demand	to demand
af-sende	Particle	V	–	send	to send off
af-spejle	Particle	V	–	mirror	to reflect
af-vente	Particle	V	–	wait	to await
af-vige	Particle	V	–	retreat	to diverge
bag-binde	Particle	V	back/behind	bind	to tie someone's hands behind his back
bag-gård	Particle	N	–	yard	backyard
bag-hjul	Particle	N	–	wheel	backwheel
bag-hoved	Particle	N	–	head	back of the head
bag-krop	Particle	N	–	body	hind part of the body
bag-linje	Particle	N	–	line	back line
bag-lomme	Particle	N	–	pocket	back pocket
bag-pote	Particle	N	–	paw	hind paw
bag-tæppe	Particle	N	–	carpet	back cloth
bag-vægt	Particle	N	–	weigth	preponderance
om-bejle	Particle	V	around/re-	court	to court
om-døbe	Particle	V	–	baptize	to rename
om-eksamen	Particle	N	–	examination	re-examination
om-kranse	Particle	V	–	wreathe	to encircle
om-kreds	Particle	N	–	circle	circumference
om-ryste	Particle	V	–	shake	to shake
om-skole	Particle	V	–	school	to retrain
om-sværme	Particle	V	–	hover	to hover around
om-vej	Particle	N	–	road	detour
op-brud	Particle	N	up	break	departure
op-digte	Particle	V	–	invent	to invent, to make up
op-dyrke	Particle	V	–	cultivate	to cultivate
op-fange	Particle	V	–	catch	to catch
op-fordre	Particle	V	–	encourage	to encourage
op-hæve	Particle	V	–	raise	to abolish
op-høje	Particle	V	–	heighten	to raise (up)
op-kalde	Particle	V	–	call	to name after
op-løse	Particle	V	–	loosen	to dissolve
op-takt	Particle	N	–	bar (music)	upbeat
til-bede	Particle	V	to/towards	pray	to worship
til-flugt	Particle	N	–	escape	refuge
til-knytte	Particle	V	–	tie	to attach
til-kæmpe	Particle	V	–	fight	to win by hard work
til-lære	Particle	V	–	learn	to learn
til-løb	Particle	N	–	run	run-up
til-råb	Particle	N	–	shout	shout
til-stræbe	Particle	V	–	strive	to aim at
til-støde	Particle	V	–	push	to befall
til-træde	Particle	V	–	step	to begin, to accept
be-fri	Prefix	V	TRANS/DIR	free	to free
be-grave	Prefix	V	–	dig	to bury
be-klage	Prefix	V	–	wail	to regret
be-laste	Prefix	V	–	load	to load (excessively)
be-mærke	Prefix	V	–	mark	to notice
be-ordre	Prefix	V	–	order	to order
be-sejre	Prefix	V	–	win	to vanquish
be-skærme	Prefix	V	–	screen	to shield
be-spotte	Prefix	V	–	ridicule	to ridicule
be-vogte	Prefix	V	–	guard	to guard
gen-bruge	Prefix	V	re	use	to reuse
gen-danne	Prefix	V	–	create	to recreate
gen-finde	Prefix	V	–	find	to find again

Table A1 (*continued*)

Word	Type	PoS	Gloss 1	Gloss 2	Translation
gen-lyd	Prefix	N	–	sound	echo
gen-læse	Prefix	V	–	read	to reread
gen-rejse	Prefix	V	–	raise	to reerect
gen-skin	Prefix	N	–	shine	reflection
gen-starte	Prefix	V	–	start	to restart
gen-syn	Prefix	N	–	sight	reunion
gen-vælge	Prefix	V	–	elect	to reelect
mis-farve	Prefix	V	mis/dis	color	to discolor
mis-greb	Prefix	N	–	grasp	mistake
mis-klang	Prefix	N	–	sound	dissonance
mis-klæde	Prefix	V	–	suit	to ill-suit
mis-røgt	Prefix	N	–	care	neglect
mis-tanke	Prefix	N	–	thought	suspicion
mis-tolke	Prefix	V	–	interpret	to misinterpret
mis-tro	Prefix	N	–	belief	mistrust
mis-vise	Prefix	V	–	show	to mislead
mis-vækst	Prefix	N	–	growth	crop failure
sam-arbejde	Prefix	N	joint	work	to collaborate
sam-drift	Prefix	N	–	operation	joint operation
sam-handel	Prefix	N	–	trade	trade
sam-køre	Prefix	V	–	drive	to coordinate
sam-liv	Prefix	N	–	life	common life
sam-råd	Prefix	N	–	council	consultation, council
sam-spil	Prefix	N	–	play	interplay
sam-tale	Prefix	N	–	speech	conversation
sam-tid	Prefix	N	–	time	contemporary time
sam-vær	Prefix	N	–	being	being together
u-fred	Prefix	N	un	peace	discord
u-held	Prefix	N	–	luck	bad luck
u-jævn	Prefix	A	–	even	uneven
u-klog	Prefix	A	–	wise	unwise
u-skarp	Prefix	A	–	sharp	blurred
u-sund	Prefix	A	–	healthy	unhealthy
u-tryg	Prefix	A	–	safe	unsafe
u-tæt	Prefix	A	–	dense	leaky
u-ven	Prefix	N	–	friend	enemy
u-ægte	Prefix	A	–	genuine	fake

Table A2
Items from Experiments 2 and 3, with information about morphological type and part of speech as well as glosses for first and second constituents and whole-word translations. For affixes, the general meaning of the affix is given once. For the prefix *be-*, Trans/dir refers to its function of making its stem transitive or adding directional meaning.

Word	Type	PoS	Gloss 1	Gloss 2	Translation
af-hente	Particle	V	off/away	fetch	to collect
af-magt	Particle	N	–	power	powerlessness
af-savn	Particle	N	–	lack	privation
af-skaffe	Particle	V	–	get	to do away with
af-sløre	Particle	V	–	veil	to unveil
bag-vende	Particle	V	back/behind	turn	to turn wrong side round
bag-klog	Particle	A	–	wise	wise after the event
bag-sæde	Particle	N	–	seat	backseat
bag-side	Particle	N	–	side	back, reverse
bag-vagt	Particle	N	–	guard	person on call
efter-gilde	Particle	N	after	party	after party
efter-løn	Particle	N	–	salary	early retirement
efter-mæle	Particle	N	–	voice	posthumous reputation
efter-smag	Particle	N	–	taste	aftertaste
efter-skælv	Particle	N	–	quake	aftershock
med-bringe	Particle	V	with	bring	to bring
med-fange	Particle	N	–	prisoner	fellow prisoner
med-vind	Particle	N	–	wind	tailwind
med-virke	Particle	V	–	work	to take part

(*continued on next page*)

Table A2 (*continued*)

Word	Type	PoS	Gloss 1	Gloss 2	Translation
med-ynk	Particle	N	–	pitiful sight	pity
om-egn	Particle	N	around/re-	area	neighborhood
om-favne	Particle	V	–	embrace	to embrace
om-rids	Particle	N	–	sketch	outline
om-serv	Particle	N	–	serve (sports)	second serve
om-verden	Particle	N	–	world	surrounding world
op-fatte	Particle	V	up	grasp	to understand
op-finde	Particle	V	–	find	to invent
op-nå	Particle	V	–	reach	to achieve
op-sving	Particle	N	–	swing	upswing
op-vask	Particle	N	–	wash	washing-up
over-moden	Particle	A	over	ripe	overripe
over-skue	Particle	V	–	see	to foresee, cope with
over-skygge	Particle	V	–	shadow	to overshadow
over-tale	Particle	V	–	speak	to persuade
over-tone	Particle	N	–	tone	overtone
til-kalde	Particle	V	to/towards	call	to call/send for
til-navn	Particle	N	–	name	byname, nickname
til-passe	Particle	V	–	fit	to adjust
til-snit	Particle	N	–	cut	form
til-trække	Particle	V	–	draw	to attract
be-koste	Prefix	V	TRANS/DIR	cost	to pay for
be-laste	Prefix	V	–	load	to load (excessively)
be-nægte	Prefix	V	–	deny	to deny
be-slutte	Prefix	V	–	conclude	to decide
be-snakke	Prefix	V	–	talk	to persuade, to coax
gen-digte	Prefix	V	re	write poetry	to retell, reproduce
gen-kende	Prefix	V	–	know	to recognize
gen-klang	Prefix	N	–	sound	echo
gen-lære	Prefix	V	–	learn	relearn
gen-spejle	Prefix	V	–	mirror	reflect
mis-kredit	Prefix	N	mis/dis	credit	discredit
mis-lyd	Prefix	N	–	sound	dissonance
mis-mod	Prefix	N	–	courage	despondency
mis-tolke	Prefix	V	–	interpret	misinterpret
mis-unde	Prefix	V	–	grant	begrudge
sam-drift	Prefix	N	joint	operation	joint operation
sam-ordne	Prefix	V	–	organize	coordinate
sam-råd	Prefix	N	–	council	consultation, council
sam-sende	Prefix	V	–	send	to broadcast simultaneously
sam-tænke	Prefix	V	–	think	integrate
u-fiks	Prefix	A	un	smart	clumsy, unattractive
u-klar	Prefix	A	–	clear	unclear
u-lykke	Prefix	N	–	happiness	accident, tragedy
u-stabil	Prefix	A	–	stable	unstable
u-vejr	Prefix	N	–	weather	storm
und-gå	Prefix	V	avoid/without	go	avoid
und-skylde	Prefix	V	–	owe	apologise
und-slippe	Prefix	V	–	slip	escape
und-være	Prefix	V	–	be	be/do without
und-vige	Prefix	V	–	yield/retreat	evade
van-ære	Prefix	N	mis/dis	honor	dishonor
van-held	Prefix	N	–	luck	bad luck
van-hellig	Prefix	N	–	holy	profane
van-røgte	Prefix	V	–	care	neglect
van-skæbne	Prefix	N	–	fate	misfortune
fabel-agtig	Suffix	A	fable	-like	phenomenal
fejl-agtig	Suffix	A	error	–	wrong
fløjls-agtig	Suffix	A	velvet	–	velvety
løgn-agtig	Suffix	A	lie	–	mendacious
tumult-agtig	Suffix	A	riot	–	tumultuous
brug-bar	Suffix	A	use	-able	useful
bær-bar	Suffix	A/N	carry	–	portable, laptop

Table A2 (*continued*)

Word	Type	PoS	Gloss 1	Gloss 2	Translation
læs-bar	Suffix	A	read	–	readable
mærk-bar	Suffix	A	feel	–	noticeable
print-bar	Suffix	A	print	–	printable
accept-ere	Suffix	V	accept	verbalising	to accept
billett-ere	Suffix	V	ticket	–	to collect tickets
march-ere	Suffix	V	march	–	to march
process-ere	Suffix	V	process	–	to process
respekt-ere	Suffix	V	respect	–	to respect
bygg-eri	Suffix	N	build	-ery	building(s)
fjoll-eri	Suffix	N	play the fool	–	nonsense
gætt-eri	Suffix	N	guess	–	guesswork
ras-eri	Suffix	N	rage	–	rage
tyv-eri	Suffix	N	thief	–	theft
frisk-hed	Suffix	N	fresh	-ness	freshness
gerrig-hed	Suffix	N	miserly	–	miserliness
korrekt-hed	Suffix	N	correct	–	correctness
tæt-hed	Suffix	N	close	–	closeness
tavs-hed	Suffix	N	silent	–	silence
asket-isk	Suffix	A	ascetic	-ic	ascetic
film-isk	Suffix	A	film	–	cinematic
gigant-isk	Suffix	A	giant	–	gigantic
idyll-isk	Suffix	A	idyl	–	idyllic
kult-isk	Suffix	A	cult	–	cultic
arve-lig	Suffix	A	inherit	-like/-able	hereditary
kede-lig	Suffix	A	bore	–	boring
natur-lig	Suffix	A	nature	–	natural
pynte-lig	Suffix	A	decorate	–	neat, decorative
sommer-lig	Suffix	A	summer	–	summerly
miljø-mæssig	Suffix	A	environment	-al/-related	environmental
motiv-mæssig	Suffix	A	motive/motif	–	with regard to motives/motifs
regel-mæssig	Suffix	A	rule	–	regular
rutine-mæssig	Suffix	A	routine	–	routine
vækst-mæssig	Suffix	A	growth	–	with regard to growth
arrig-skab	Suffix	N	bad-tempered	-ness/-ship	bad temper
doven-skab	Suffix	N	lazy	–	laziness
mester-skab	Suffix	N	master	–	championship
svanger-skab	Suffix	N	pregnant	–	pregnancy
ven-skab	Suffix	N	friend	–	friendship
føl-som	Suffix	A	feel	-ful	sensitive
glem-som	Suffix	A	forget	–	forgetful
gru-som	Suffix	A	horror	–	terrible
skån-som	Suffix	A	spare	–	protective
stræb-som	Suffix	A	strive	–	hardworking

References

Baayen, R. (2010). The directed compound graph of English: An exploration of lexical connectivity and its processing consequences. In S. Olsen (Ed.), *New impulses in word-formation (Linguistische Berichte Sonderheft 17)* (pp. 383–402). Hamburg: Buske.

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

Baayen, R., Dijkstra, T., & Schreuder, R. (1997). and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language, 36*, 94–117.

Baayen, R., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 53*, 496–512.

Baayen, R., McQueen, J., Dijkstra, T., & Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In R. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing*. Berlin: Mouton de Gruyter, pp. 355–390.

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

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (cd-rom). University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baayen, R., & Schreuder, R. (1999). War and peace: Morphemes and full forms in a non-interactive activation parallel dual route model. *Brain and Language, 68*, 27–32.

Baayen, R., & Schreuder, R. (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences), 358*, 1–13.

Baayen, R., Wurm, L., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon, 2*, 419–463.

Balling, L. (2008). Morphological effects in Danish Auditory Word Recognition (PhD thesis). Aarhus: University of Aarhus.

Balling, L., & Baayen, R. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes, 23*, 1159–1190.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 340–357.

Bertram, R., Laine, M., Baayen, R., Schreuder, R., & Hyönä, J. (1999). Affixal homonymy triggers full-form storage even with inflected words, even in a morphologically rich language. *Cognition, 74*, B13–B25.

Bertram, R., Laine, M., & Karvinen, K. (1999). The interplay of word formation type, affixal homonymy, and productivity in lexical processing: Evidence from a morphologically rich language. *Journal of Psycholinguistic Research, 100*, 213–226.

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading diffculty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*, 1–12.

Butterworth, B. (1983). Lexical representation. In *Language production: Development*. In B. Butterworth (Ed.). *Writing and other language processes* (Vol. II). London: Academic Press, pp. 257–294.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–258.

Crawley, M. J. (2002). *Statistical computing. An introduction to data analysis using S-plus*. Chichester: Wiley.

Cutler, A. (1981). Degrees of transparancy in word formation. *CLS, 26*, 73–77.

Davis, M., Marslen-Wilson, W., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 218–244.

De Jong, N. H., Feldman, L. B., Schreuder, R., Pastizzo, M., & Baayen, R. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. *Brain and Language, 81*, 555–567.

De Jong, N. H., Schreuder, R., & Baayen, R. (2000). The morphological family size effect and morphology. *Language and Cognitive Processes, 15*, 329–365.

De Vaan, L., Schreuder, R., & Baayen, R. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon, 2*, 1–23.

Forster, K. I., & Forster, J. (2006). DMDX version 3.1.4.5. <http://www.u.arizona.edu/jforster/dmdx.htm> Accessed January 2006.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1139–1144).

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl '02)*. Ann Arbor: Association for Computational Linguistics.

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse tree of sentences as a function of the sentence number. In *Proceedings of the conference on empirical methods in natural language processing*. Sapporo.

Hale, J. (2001). A Probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the asssociation for computational linguistics*.

Hansen, A. (1967). *Moderne Dansk*. Copenhagen: Grafisk Forlag.

Hay, J. B. (2002). From speech perception to morphology: Affix-ordering revisited. *Language, 78*, 527–555.

Janssen, N., Bi, Y., & Caramazza, A. (2008). A tale of two frequencies: Determining the speed of lexical access for Mandarin Chinese and English compounds. *Language and Cognitive Processes, 23*(7), 1191–1223.

Kemps, R., Ernestus, M., Schreuder, R., & Baayen, R. (2005). Prosodic cues for morphological complexity: The case of Dutch noun plurals. *Memory and Cognition, 33*, 430–446.

Kuperman, V., Bertram, R., & Baayen, R. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language, 62*, 83–97.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP, 35*, 876–895.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

Libben, G., Gibson, M., Yoon, Y., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language, 84*, 50–64.

Lukatela, G., Eaton, T., Lee, C., Carello, C., & Turvey, M. (2002). Equal homophonic priming with words and pseudohomophones. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 3–21.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 1–24.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial overview. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance x: Control of language processes* (pp. 125–150). Hillsdale: Erlbaum.

Marslen-Wilson, W. D., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101*, 3–33.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.

Meunier, F., & Segui, J. (1999). Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language, 41*, 327–344.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*, 1271–1278.

Moscoso del Prado Martín, F., Deutsch, A., Frost, R., Schreuder, R., De Jong, N. H., & Baayen, R. (2005). Changing places: A cross-language perspective on frequency and family size in Hebrew and Dutch. *Journal of Memory and Language, 53*, 496–512.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*, 189–234.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*(2), 327–357.

Norris, D., & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395.

Norris, D., McQueen, J., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology* (3), 191–243.

R Development Core Team (2011). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria. <http://www.R-project.org/>. ISBN: 3-900051-07-0.

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes* (7), 942–971.

Salverda, A., Dahan, D., & McQueen, J. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51–89.

Schreuder, R., & Baayen, R. (1997). How complex simplex words can be. *Journal of Memory and Language, 37*, 118–139.

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In M. Coltheart (Ed.), *Attention and Performance XII*. Hove: Lawrence Erlbaum Associates, pp. 245–264.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116*, 71–86.

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from either…or. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 425–436.

Tabak, W., Schreuder, R., & Baayen, R. (2005). Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. In S. Kepser & M. Reis (Eds.), *Linguistic evidence—empirical. Theoretical, and computational perspectives* (pp. 529–555). Berlin: Mouton de Gruyter.

Tabak, W., Schreuder, R., & Baayen, R. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon, 5*(1), 22–46.

Van Casteren, M. (2006). Mix. <http://www.mrc-cbu.cam.ac.uk/maarten/Mix.htm> Accessed January 2006.

Wood, S. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear. *Journal of the Royal Statistical Society (B), 73*, 3–36.

Wurm, L. (1997). Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language, 37*, 438–461.

Wurm, L., Ernestus, M., Schreuder, R., & Baayen, R. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and conditional root uniqueness points. *The Mental Lexicon*, 125–146.

Wurm, L., & Ross, S. E. (2001). Conditional root uniqueness points: Psychological validity and perceptual consequences. *Journal of Memory and Language, 45*, 39–57.