Surviving blind decomposition: a distributional analysis of the time-course of complex word recognition

Daniel Schmidtke
McMaster University, Canada

Kazunaga Matsuki
McMaster University, Canada

Victor Kuperman
McMaster University, Canada

March 14, 2017

Running Head: Surviving blind decomposition

Corresponding author:
Daniel Schmidtke
University of Alberta
Department of Psychology
Biological Sciences Building
Edmonton, Alberta
Canada T6G 2R3
Email: schmidtk@ualberta.ca

Abstract

The current study addresses a discrepancy in the psycholinguistic literature about the chronology of information processing during the visual recognition of morphologically complex words. *Form-then-meaning* accounts of complex word recognition claim that morphemes are processed as units of form prior to any influence of their meanings, whereas *form-and-meaning* models posit that recognition of complex word forms involves the simultaneous access of morphological and semantic information. The study reported here addresses this theoretical discrepancy by applying a non-parametric distributional technique of survival analysis (Reingold & Sheridan, 2014) to two behavioural measures of complex word processing. Across seven experiments reported here, this technique is employed to estimate the point in time at which orthographic, morphological and semantic variables exert their earliest discernible influence on lexical decision reaction times and eye movement fixation durations. Contrary to form-then-meaning predictions, Experiments 1-4 reveal that surface frequency is the earliest lexical variable to exert a demonstrable influence on lexical decision reaction times for English and Dutch derived words (e.g., *badness*; *bad* + -*ness*), English pseudo-derived words (e.g., *wander*; *wand* + -*er*) and morphologically simple control words (e.g., *ballad*; *ball* + -*ad*). Furthermore, for derived word processing across lexical decision and eye-tracking paradigms (Experiments 1-2; 5-7), semantic effects emerge early in the time-course of word recognition, and their effects either precede or emerge simultaneously with morphological effects. These results are not consistent with the premises of the form-then-meaning view of complex word recognition, but are convergent with a form-and-meaning account of complex word recognition.

Keywords: morphological processing; survival analysis; lexical decision; semantics; valence

Acknowledgments This work was supported by the Ontario Trillium Award and a Graduate fellowship awarded by the Lewis & Ruth Sherman Centre for Digital Scholarship (McMaster University) to the first author. Daniel Schmidtke is now at the Department of Psychology at the University of Alberta, Canada.

The second and third authors' contributions were supported by the NIH R01 HD 073288 (PI Julie A. Van Dyke) grant for the project *Retrieval interference in skilled and unskilled reading comprehension.* The third author's contribution was also supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant 402395-2012, the SSHRC Partnership Training grant 895-2016-1008 (PI Gary Libben), the Early Research Award from the Ontario Ministry of Research and Innovation, and the Canada Research Chair (Tier 2) award.

We are indebted to Harald Baayen for his valuable feedback, comments and discussion at the early stages of this work. We also thank the audience at the Tenth International Conference on The Mental Lexicon, Ottawa, ON, Canada (October 19th - 21st). Thanks are also due to Thomas Spalding, Christina Gagné and members of the Complex Cognition Lab at the University of Alberta for feedback and discussion on an early presentation of this work.

Surviving blind decomposition:

a distributional analysis of the time-course of complex word recognition

The question of how proficient readers recognize morphologically complex words in an apparently automatized and seamless manner has given rise to decades worth of psycholinguistic research (for a review, see Amenta & Crepaldi, 2012; Diependaele, Grainger, & Sandra, 2012). Despite advancements in the experimental and statistical resources available to psycholinguists, crucial components of the cognitive processes involved in complex word reading have been difficult to establish and sequence. This paper addresses a long-standing debate regarding the hypothesized stages during the time-course of word identification at which orthographic forms and meanings influence complex word recognition.

A dominant account of morphological processing is that the morpho-orthographic units of a complex word are analyzed by the language processing system before access to the word's meaning is able to proceed. Under this *form-then-meaning* account of complex word recognition (also known as *early morpho-orthographic decomposition* or *late-lexical-access*), complex words undergo obligatory orthographic segmentation into morphemic units in a serialised and semantics-blind manner (cf. Meunier & Longtin, 2007; Rastle & Davis, 2008; Solomyak & Marantz, 2010; Taft & Forster, 1976; Taft, 2004). According to this account, upon visual presentation of a complex word (e.g. *joker*), the word string is first decomposed into separate morphemes, i.e. *joke* + *er*. The decomposed orthographic form of the complex word then grants access to morphemic unit representations. Once the orthographic cues to the morpheme representations have been isolated, the morphemic representations are then extracted from the mental lexicon. This stage is referred to as either 'look-up' (i.e. the cognitive process of 'looking up' the component morphemes of complex word in the lexicon; Fruchter & Marantz, 2015) or 'licensing' (Meunier & Longtin, 2007). According to Meunier and Longtin (2007; pp. 469), the process of licensing "checks the appropriateness of morpheme combinations, for instance by assessing whether representations can be integrated on the basis of their subcategorization properties". At the final stage, the component morphemes of the complex word are recombined and the word meaning - e.g., 'a person who likes telling jokes' - is obtained. One corollary of this theoretical view is that any word with apparent morphological structure (e.g., *corner*) undergoes the same sequence of processing stages as true complex words like *joker*. In sum, under this processing theory, morpho-orthographic access is obligatory, is encapsulated from and must precede semantic access of the whole word.

Evidence from masked-priming research

There exists plentiful evidence in support of obligatory early morpho-orthographic decomposition among healthy adult populations. A main source of this evidence comes from the results of forward masked priming studies (for a review see Rastle and Davis, 2008). The masked priming paradigm is employed under the assumption that a briefly presented prime before the presentation of a target item, with a stimulus onset asynchrony (SOA) of < 60 ms, exploits unconscious early cognitive processing (Forster, Mohan, & Hector, 2003). Most priming studies find no reliable difference in the amount of priming between a condition in which a prime is morphologically and semantically related to a simplex target word (i.e. semantically transparent; *trucker-TRUCK*) and a condition in which the prime only appears to be morphologically related but is not semantically related to the target (i.e. semantically opaque; *corner-CORN*[1]). This null effect of semantic transparency, relative to an orthographic control condition (i.e. no morphological relation between prime and target; *brothel-BROTH*) has been replicated across multiple experiments and languages, including English (e.g., Rastle, Davis, & New, 2004; Marslen-Wilson, Bozic, & Randall, 2008; Beyersmann, Ziegler, Castles, Colheart, Kezilas, & Grainger, 2015), French (Longtin, Segui, & Hallé, 2003), Spanish (Lazaro, Illera, & Sainz, 2016), Korean (Kim, Wang, & Taft, 2015) and Russian (Kazanina, 2011). This pattern of results is taken to suggest that the semantics of either the entire prime (e.g., *trucker*) or that of its base morpheme (e.g., *truck*) is not accessed within the timeframe set by the duration of the SOA. The finding, i.e. statistically equivalent priming magnitudes across pseudo and true derived words, is interpreted as support for semantics-blind morpho-orthographic segmentation.

The aforementioned body of evidence has been challenged by a growing number of studies implementing either masked priming or paradigms with very short SOAs. These studies demonstrate reliable semantic priming effects in Dutch, English, Finnish and Spanish (for extensive discussions, see Davis & Rastle, 2010; Feldman, O'Connor, & Moscoso del Prado Martín, 2009; Feldman, Milin, Cho, Moscoso del Prado Martín, & O'Connor, 2015; Van den Bussche, Van den Noorgate, & Reyn-

[1] This condition is labelled as semantically opaque relative to the transparent condition. However, as pointed out by Feldman, Milin, Cho, Moscoso del Prado Martín, and O'Connor (2015), many of the target stimuli used in this condition (see e.g., Rastle, Davis, & New, 2004) do not function as a true morphological stem. For example, *corn* does not serve as a true morpheme in *corner*, even though the latter includes a true suffix, *-er*. We therefore treat these prime-target pairs as a 'pseudo-suffixed' condition and the transparent prime-target pairs as a 'true derived' condition.

voet, 2009). For instance, Feldman et al. (2015) used a within-experiment manipulation, which included exposing participants to five different SOAs all below 100 ms: they observed shorter latencies to the semantically related conditions in very short SOAs of 34 and 48 ms. Semantic effects also emerged in a cross-case 'same-different' task by Duñabeitia, Kinoshita, Carreiras, and Norris (2011). This task requires participants to assess the likeness of a lowercase and an uppercase letter string presented sequentially. They found no difference in priming magnitude across conditions containing true suffix strings (e.g., *trucker-TRUCK* and *corner-CORN*) and orthographic controls (e.g., *brothel-BROTH*), suggesting that semantics-blind morpho-orthographic segmentation is not obligatory. This set of results is compatible with a host of theories of complex processing, that while varying in precise details, assume that morphological activation is conditioned simultaneously by form and meaning characteristics of the complex word (Baayen & Schreuder, 1999; Diependaele, Sandra, & Grainger, 2005; Diependaele, Duñabeitia, Morris, & Keuleers, 2011; Grainger & Ziegler, 2001; Kuperman, Bertram, Schreuder, & Baayen, 2009; Libben, 2005, 2014; Moscoso del Prado Martín, 2007), or that the activation of complex word meaning is not aided by a morphological level of representation at all (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Plaut & Gonnerman, 2000).

The debate on the timing of semantic access is difficult to resolve within the masked priming technique, because – as acknowledged by Rastle and Davis (2008, pp. 949) – it may be "impossible" to infer the time-course of word recognition from a prime duration. An SOA of 40 ms does not indicate that decomposition occurs at 40 ms after presentation of the visual stimulus, and so masked priming with short SOAs is a suboptimal technique for characterizing the absolute timepoint at which this cognitive process occurs. This lack of temporal precision can be remedied either by the use of experimental paradigms with a fine temporal resolution, or statistical methods that enable early detection of an effect in the neuro/behavioural signal, or both. In the following two subsections of the Introduction we present results obtained from these paradigms, while the final subsection discusses a statistical method for detecting onsets of effects, i.e. survival analysis.

Evidence from neurophysiological studies

Magneto- and electroencephalography (MEG and EEG) allow for an excellent temporal resolution of neural behaviour as it unfolds during visual word recognition. Results of cross-linguistic MEG and EEG recorded during primed or unprimed lexical decision on derived, inflected and compound words serve as a second source of empirical support for the form-then-meaning accounts of morphological processing. A number of experiments in English and French (e.g., Lavric, Clapp, & Rastle, 2007; Lavric, Elchlepp, & Rastle, 2012; Morris, Frank, Grainger, & Holcomb, 2007; Royle, Drury, Bourguignon, & Steinhauer, 2010; 2012; Morris & Stockall, 2012) combined event-related brain potentials (ERPs) with lexical decision tasks and reported that early components of neural activity are devoted to the processing of morpho-orthographic features. These studies reported that neural activity at 100-200 ms post-onset of the visual stimulus is only sensitive to visual features of the prime and the target, while semantic relatedness of the prime and target affected later neural activity, between 300-500 ms post-onset of the visual stimulus. This time-course is in line with the premises of the form-then-meaning account which was developed on the basis of behavioural results (see above section).

To take an example of one such experiment, in an unprimed lexical decision-and-ERP experiment, Lavric, Elchlepp, and Rastle (2012) reported an orthographic correlate of morphological decomposition taking place about 190 ms post-onset in the ERP signal, while differences between processing semantically transparent derived words (e.g., *darkness*) and pseudo-derived words (e.g., *corner*) were detected in neural activity approximately 70 ms later, around 260 ms post-onset (see Zweig and Pylkkänen, 2009, for converging MEG data). In addition, Lavric, Rastle, and Clapp (2011) reported a masked priming lexical decision-and-ERP study with long SOAs which contrasted true derived prime-target pairs (e.g., *magical-MAGIC*), pseudo-suffixed pairs (e.g., *compassion-COMPASS*) and orthographic control pairs (e.g., *brothel-BROTH*). The authors found effects of semantic transparency (as gauged by Latent Semantic Analysis; Landauer & Dumais, 1997) at 400 ms post stimulus onset (a late N400 component). Based on these results, the authors concluded that, "[t]he lateness of the effects of semantic transparency favours a single, orthography-based, mechanism of morphological decomposition 'licensed' at a late processing stage" (Lavric et al., 2011, pp. 684).

In addition, several MEG studies also report that effects of any and all semantic properties on lexical decision performance are only detectable in late components of complex word recognition. Specifically, MEG studies appear to converge on the M350 component (defined as a peak in neural activity approximately 350 ms after stimulus presentation) as the processing time window that is sensitive to lexical frequency, semantic transparency, morphological family frequency and size (i.e. the number and cumulative frequency of complex words sharing a morpheme), and lexical ambiguity (e.g., homonymy, heteronomy and polysemy), see Beretta, Fiorentino, and Poeppel (2005), Simon, Lewis, and Marantz (2012), Solomyak and Marantz (2009, 2010), and Lewis, Solomyak, and Marantz (2011); and also reviews in Fiorentino and Poeppel (2007) and Pylkkänen, Llinás, and Murphy (2006).

Of particular relevance to the current investigation, Solomyak and Marantz (2010) carried out a MEG study which examined neural sensitivity to statistical properties in a sequence of time intervals. Solomyak and Marantz (2010) reported initial brain activity associated with orthography-specific properties of derived words (M130 component), followed by lexical properties indexing parsing of the stem and affix (M170 component), and a final "lemma" activation stage (M350) as indicated by an increase in neural activity associated with word frequency. The results of this study are well-suited to the time-course propounded in the semantics-blind obligatory morphological decomposition account, a prediction which is formalized by Solomyak and Marantz (2010) as follows:

> 1. In the first stage of processing, potential affixes are recognized by form, and parsing between stem and affix is attempted based on activation of the (visual) word form of the stem. Thus, we expect affix-specific variables such as affix frequency to correlate first with brain activity, followed by variables associated with parsing, such as the transition probability between stem and affix. Parsing of stem and affix leads to lexical access for the stem. At this stage, lexical variables including lemma frequency should be relevant, even for bound roots... Following lexical access to the stem, recombination of stem and affix occurs. Here, variables such as the transition probability between affix and stem and/or surface frequency should correlate with brain activity. (Solomyak & Marantz, 2010, pp. 2045)

With a similar set of predictions, Fruchter and Marantz (2015) recently investigated the time-course of English word recognition in a MEG experiment involving lexical decision to derived words. They reported left-temporal activity associated with derivational family entropy as early as 240 ms, followed by a facilitatory effect of surface frequency, in the same region, at a later time window of 430-500 ms. They also reported an effect of semantic coherence of the derived word at the left orbitofrontal region of interest, at 300-500 ms. The results of Fruchter and Marantz

(2015) confirmed their prediction that "semantic effects, like the effects of surface frequency, should occur after the effects of derivational family entropy, which relate to stem lookup" (pp. 82). In addition, the authors argued that these results are compatible with the timing of effects outlined in the obligatory decomposition model, in which base (stem) frequency effects "always" emerge at the early decomposition stage, while "surface frequency has its impact at the subsequent combination stage" (Taft, 2004, pp. 762). Moreover a study of morphological processing of inflected forms in Finnish (Lehtonen, Cunillera, Rodríguez-Fornells, Hultén, Tuomainen, & Laine, 2007) found late main effects of morphological complexity (relative to monomorphemic words) emerging in the average ERP waveform, at between 550 and 850 ms. The authors argue that while interactive effects of word frequency by morphology emerged earlier (450-550 ms), reflecting access to full form representations, the late effect of morphological complexity reflects a post-access semantic-syntactic level of processing.

It is worth noting that a separate body of ERP and MEG research demonstrates much earlier (100-200 ms post-onset) effects of word frequency, lexicality, semantic coherence of a word's morphological family and other semantic lexical properties in the time course of recognizing morphologically simple words (Assadollahi & Pulvermüller, 2003; Hauk, Davis, Ford, Pulvermüller, & Marslen-Wilson, 2006; Sereno, Rayner, & Posner, 1998; Pulvermüller, 2002; Penolazzi, Hauk, & Pulvermüller, 2007; Reichle, Tokowicz, Liu, & Perfetti, 2011). Also, in a recent ERP study, Jared, Jouravlev and Joanisse (2016) found effects of semantic transparency during visual English derived word recognition in a time window as early as 200-250 ms. Taken together, these findings substantially shorten the time window for the access of word meaning as compared to the neuroimaging studies surveyed above.

To sum up, the fine temporal resolution afforded by MEG and ERP techniques favours the outline of form-then-meaning theories of complex word recognition. That is, reported cortical activity associated with the processing of printed complex words suggests that orthographic segmentation into morphemic units is obligatory and proceeds relatively early during word processing in a strictly serial and semantics-blind manner. Critically, we argue that this largely coherent and copious body of evidence does not produce a plausible timeline of complex word processing. Specifically, this evidence is logically incompatible with time signatures of effects elicited by semantic and formal lexical properties in behavioral studies of word recognition. The following section elaborates on this issue.

Behavioral effects precede neurophysiological effects

Brain activation in response to a word property must be a cause of any change in behavior that this property might elicit. This makes the time signature of an effect in the behavioral record the temporal upper bound for that effect's manifestation in the neural record. Even more precisely, stimulus-related brain activity is expected to sufficiently inform a behavioral response before the program of that (manual, digital, oculomotor or articulatory) response reaches a non-labile, non-cancelable stage of execution: the upper bound is thus the logical initiation of said behavioral stage.

It is fairly easy to demonstrate that the time windows reported by neurophysiological studies of complex words (reviewed above) largely violate respective behavioral upper bounds, creating a paradox wherein a behavioral effect apparently precedes the brain activity that leads to it. First, consider results of a behavioral paradigm with a high temporal resolution, i.e. eye tracking. An average fixation time on a word presented in context is in the 200-250 ms range: this estimate is true of derived words as well (cf. among others Kuperman, Bertram, & Baayen, 2010). A very similar estimate of average fixation durations, below 250 ms, holds for words presented in isolation too (O'Regan & Levy-Schoen, 1987; Vitu, 1991; Vitu, Oregan, & Mittau, 1990; see also Kuperman et al., 2009; Marelli, Amenta, Morone, and Crepaldi, 2013, Experiment 1; Miwa, Libben, Dijkstra, & Baayen, 2014; Miwa & Dijkstra, 2016). Moreover, most words – in context or isolation – are fixated only once (see Rayner, 1998). This single fixation duration shows sensitivity to a host of lexical-semantic properties of a word, such as its dominant and subordinate meanings, its emotional connotations, imageability, concreteness, homonymy and polysemy (Rayner, 1998; Staub & Rayner, 2007; Schotter & Rayner, 2012). For derived word reading, single fixation durations also show effects of surface frequency (e.g., frequency of the whole form *trucker*), stem frequency (e.g., frequency of *truck*), as well as semantic properties of the stem and the suffix (Kuperman et al., 2010 and references therein). The finding that the behaviour of the visuo-oculomotor system is influenced by complex word meaning within 200-250 ms of the stimulus fixation is clearly incompatible with either 260, 350 or 400 ms proposed as loci of semantic effects in much of the MEG and EEG literature on complex word processing. This behavior however is fully in line with the 140-200 ms locus offered by brain imaging research on simplex words (see above).

Second, eye tracking data runs counter to some of the specific temporal estimates of lexical effects suggested in the neuroimaging literature. Consider the effect of surface (whole word) frequency. The form-then-meaning account predicts it to be late (Fruchter and Marantz, 2015; Taft, 1979, among others) in the timeline of complex word processing. Yet, oculomotor evidence shows that timing of the surface frequency effect is similar for complex and simplex words, and that the effect emerges much earlier than estimated in neurophysiological studies. For instance, Pollatsek, Hyönä, and Bertram (2000; experiment 2) found sizable effects of surface frequency on first fixation durations of both Finnish compounds and matched monomorphemic words (average first fixation duration across compounds = 198 ms, and monomorphemic words = 205 ms). This finding indicates that the influence of surface frequency is not delayed for compounds relative to the monomorphemes and emerges no later than 200 ms post-onset. For similar findings in English derived words, see Kuperman and Van Dyke (2011a). Furthermore, in a recent series of non-parametric distributional analyses, Staub, White, Drieghe, Hollway, and Rayner (2010) re-analysed data obtained from two eye movement experiments in which participants silently read critical (simplex and complex) words that were embedded within sentences (Drieghe, Rayner, & Pollatsek, 2008; White, 2008). Using the vincentile plotting technique (Ratcliff, 1979; Vincent, 1912), Staub et al. (2010) showed that word frequency exerts an influence on eye movement fixations (for first fixation duration and gaze duration) as early as 180-200 ms. In sum, these and other results (e.g., Rayner, Liversedge, White, & Vergilino-Perez, 2003; Rayner, Ashby, Pollatsek, & Reichle, 2004) demonstrate that the word frequency effect influences the very initial stages of lexical processing during reading and that the timing of the effect may be indifferent to a word's morphological complexity.

Third, eye tracking studies have revealed that lexical and semantic factors reliably affect the earliest eye movements to Dutch, English, Finnish and Italian derived and compound words (Juhasz & Berkowitz, 2011; Kuperman, Bertram, & Baayen, 2008; Kuperman, Schreuder, Bertram, & Baayen; 2009; Marelli and Luzzatti, 2012)[2]. For example, Kuperman et al. (2009) registered eye movements during a lexical decision task in which compounds were presented in isolation. They found first fixation durations (average = 270 ms) on long (8-12 letter) Dutch compounds to be affected by lexical frequency of the compound as well as by the family size of the left constituent

[2] Data on compounds is relevant for the present discussion, since the form-then-meaning account predicts that these words – much like derived forms – are decomposed into morphemes in a fast and obligatory fashion.

(*car* in *carwash*). Similar effects of compound frequency and left constituent family size were detected during first fixation durations (average = 221 ms) on 10-18 letter-long Finnish compounds (Kuperman et al., 2008). Moreover, Juhasz and Berkowitz (2011) observed effects of left constituent family size on the likelihood of refixation on English compound words and on gaze durations (average gaze duration average across small and large family size conditions = 272 ms). This result indicates that morphological family size influences the decision to refixate, which is a decision takes place during the time at which the first fixation on the word is made. Importantly, the length of words used in experiments listed above makes their inspection in a single fixation problematic due to limitations of visual acuity: most of the words were processed in multiple fixations. Effects of word frequency and morphological family size on early temporal measures (first fixation duration and refixation probability) indicate that semantic information is recruited in word recognition as soon as partial orthographic information becomes available, and likely before the entire word can be inspected and form-based decomposition completed (see Kuperman et al., 2008). In a similar vein, early effects of base word family size (e.g., *dream* in *dreamful*) and suffix productivity influenced single fixation durations (average duration = 245 ms) on somewhat shorter Dutch suffixed words (Kuperman, Bertram, & Baayen, 2010). Taken together, these results are difficult to reconcile with premises of the form-then-meaning approach and with the neurophysiological evidence that forms its basis.

Lastly, numerous recent studies have targeted semantic effects on complex words directly, and have found that they arise early in the eye movement record. For example, Amenta, Marelli, and Crepaldi (2015) manipulated the semantic transparency of Italian derived words which were embedded in sentences and read silently, and reported an effect of semantic transparency and stem frequency on first fixation durations (average duration = 243 ms). Amenta et al. (2015, pp. 1590) argue that their results indicate that early processing of morphologically complex words involves access to *form-and-meaning.* That is, morphological decomposition may involve concurrent access to morphemic and whole-word meanings. Furthermore, Marelli and Luzzatti (2012) manipulated both the semantic transparency and the head position in Italian compounds. They reported a reliable influence of semantic transparency as early as 231 ms for first fixation durations averaged over quartiles of their semantic transparency measure. The data presented in Marelli and Luzzatti's (2012) study serve as evidence that access to the meaning of a complex word does not wait for complete access to the word's orthography during naturalistic word reading. Finally, Marelli, Amenta, Morone, and Crepaldi (2013; experiment 1) found further eye movement effects that challenge the obligatory composition accounts. In a lexical decision experiment that measured eye movements as a dependent variable (instead of a button press latency), Marelli et al. reported an effect of semantic transparency in first fixation duration (average = 256 ms) on Italian derived words presented in isolation: as in studies reviewed above, the absence of context did not delay nor advance the emergence of semantic effects.

Evidence outlined in this section reveals a wealth of lexical and semantic effects on early eye movement measures, observed across languages and word presentation formats. It demonstrates that lexico-semantic properties such as surface word frequency, family size of constituent morphemes and semantic transparency reliably affect complex word recognition within 200-250 ms post fixation of the visual stimulus, that is, within an average duration of the first-of-many or single fixation reported in the studies cited above. In fact, even this more stringent estimate is an overestimation of the upper temporal bound for semantic effects. As argued by Dimigen et al. (2011), in order for semantic processing to influence a first fixation duration (or refixation likelihood), it must do so before saccadic motor programming enters the non-labile stage, that is, at an estimated 80 ms before the end of that fixation and the execution of the next saccade (Becker, 1991; Findlay & Harris, 1984). These observations put the behavioral upper bound for semantic effects on complex word recognition at about 120-170 ms. The temporal windows that some neurophysiological reports allocate to semantic effects (350-400 or 400-600 ms, see previous section) violate this upper bound.

It is noteworthy that the apparent paradox of a behavioral signature predating a neural signature of a specific effect is not new to the literature that considers these signatures jointly (Dambacher, Kliegl, Hofmann, & Jacobs 2006; Dambacher & Kliegl, 2007; Rayner & Clifton, 2009; Sereno et al., 1998; Sereno & Rayner, 2003). Recent co-registration studies of ERP and eye movements (Dimigen et al., 2011; Kretzschmar, Schlesewsky, & Staub, 2015) robustly report an earlier emergence of a word's contextual predictability effect on eye movements (first fixation duration, average = 234 ms) than on brain activity (N400 peak latency = 384 ms; statistical onset of N400 latency = 248 ms). That is, at the timepoint when the peak amplitude of the predictability effect was reached in the ERP signal, 96% of fixations on the target word had already terminated; and 53% of fixations on the word had terminated before the onset of the N400 wave component. As stated by Dimigen et al. (2011, p. 565), “the present data make it hard to conceive the measurable neural effects of predictability as being causal in some way for the emergence of behavioral effects, because the bulk of the predictability effects in ERPs only followed those in behavior”. Further to this, Kretzschmar et al. (2015) confirm the temporal discrepancy of the predictability effect and reveal an even more drastic one: their manipulation of word frequency produces an expected effect on eye movement latencies but not on the amplitude or latency of the N400 (or any other) wave component. This runs counter to the assumption that any behavioral change can only arise as an effect of causal neural activity. The reasons for divergent timelines and for apparent violation of causality are presently unclear, and hint at a need for more sensitive statistical analyses of brain imaging data.

In sum, our review shows that the absolute timeline originating from current brain imaging research of morphologically complex words (and, independently, of word processing in sentence context, Dimigen et al., 2011; Kretzschmar et al., 2015) is incompatible with the constraints that behavioral data imposes on the time-course of semantic effects during the recognition of complex words either in isolation or in context. This discrepancy also calls into question the validity of the relative order of formal and semantic effects proposed by the form-then-meaning account, see quotation 1. above. The next section describes the approach that we take in order to address this issue.

The current study

Our goal is to provide further insight into the temporal sequence of cognitive processes during recognition of printed derived words (e.g., *trucker*, *happiness*). We employ a relatively novel non-parametric distributional ‘survival’ approach to the analysis of reaction times (Reingold, Reichle, Glaholt, & Sheridan, 2012; Reingold & Sheridan, 2014). Distributional analyses (Balota & Abrams, 1995; Van Zandt, 2002; Vincent, 1912) and other variations on survival analysis (e.g., Rueckl & Galantucci’s, 2005 study of morphological processing) have been used previously as a valuable complement to more commonly used analyses of variance. Their strength is in the consideration of an entire distribution of response latencies, rather than just central tendencies. This enables distributional analyses to establish, for instance, whether manipulation of a target variable shifts the entire distribution by inflating all latencies, or specifically influences fast or slow responses, leading to important insights into the time-course of processing. The specific approach adopted here, the *survival analysis* technique, has been developed with the purpose of establishing the earliest point in time at which a given variable exerts an influence over a chronometric response latency. The technique estimates a survival curve for two conditions formed by levels of the target variable, and establishes the point in time at which the two survival curves for each condition begin to diverge from one another. This divergence point is taken as an estimate of the point in time at which the given variable initiates a discernible impact on response latencies (see Methods for full description).

We aim at estimating divergence timepoints for a range of predictors, many of which have been used in prior studies on morphological processing effort (e.g., Fruchter and Marantz, 2015; Moscoso del Prado Martín, Kostić, and Baayen, 2004) and are considered to be diagnostic of the processes involved in morphological decomposition. The divergence points will allow us to eschew the practice of using mean behavioral latencies as temporal upper bounds for the timeline of related neurophysiological acitivity (see previous section), and use more precise estimates of the earliest impact instead. While we question the plausibility of the absolute timeline proposed by brain-imaging studies on derived word processing, it is logically possible that the *relative* order of formal and semantic effects that the form-then-meaning account advocates is still valid, even the survival analysis restricts the loci of effects to a much shorter timescale. Thus, our study also aims to monitor the relative order in which formal and semantic effects will appear in the behavioral timeline, as (partial) evidence in favor of either the form-then-meaning or form-and-meaning accounts of morphological processing.

Estimated onsets of effects on word recognition behavior can be used directly for validating the timeline of processing proposed on the basis of prior behavioral studies including primed and unprimed lexical decision (see above). Yet it is important to realize what the causal link between neural and behavioral effects implies, and how estimates of survival analysis can relate to neurophysiological data. As argued above, with brain activity being a necessary cause of a behavior, any behavioral signature in the time-course of processing must be an upper temporal bound for related neural activation. That is, if a contrast in, say, surface frequency comes with a divergence in survival curves estimated for behavioral latencies to higher and lower frequency words, one expects this contrast to also elicit a difference in neurophysiological responses to those words, and this difference to emerge at some time prior to the behavioral divergence point. That is, an early behavioral effect needs to have a matching signature in the brain activity equally or more early, but a late behavioral effect can have an early or late counterpart in the neural signal. Finally, the presence of an effect in the ERP or MEG signal does not necessitate the presence of the same effect in the behavioral record. In our overview of predictions below, we make a simplifying assumption that the behavioral timeline reflects – by and large – the timeline of brain activity. That is, a relatively late behavioral effect is also relatively late to emerge in the neural record, and an early behavior affects the brain early too. We revise the validity of this assumption in the General Discussion.

The chronology of lexical effects in the current survival technique implementation is operationalized here as the order of divergence point estimates for each lexical property across each word condition (derived, pseudo-derived, form control). Under the predictions of the semantics-blind fast obligatory decomposition hypothesis, one would expect that variables associated with the form and meaning of the whole derived word (e.g., surface frequency, semantic transparency and emotional valence; see below for definitions) to have divergence point estimates that occur later than those associated with morphological processing. More specifically, as Solomyak and Marantz (2010) argue in quote (1), affix-specific variables, such as affix frequency, are expected to emerge first, followed by variables related to parsing of the stem and affix, such as transition probability between the stem and affix. The initial stages of recognition are also expected to be dominated by variables indicating access to orthographic form (such as orthographic neighbourhood density; Fruchter & Marantz, 2015). The influence of these properties would correspond to the initial stages of recognition, which is devoted to morpho-orthographic properties of the input. Once the morphological properties of the affix has been accessed, lexical access to the stem is expected to commence. The initiation of this process would come with the onset of the effect of lemma frequency, stem frequency or derivational family entropy (Fruchter & Marantz, 2015, pp. 83). Following this, recombination of the stem and affix is predicted to occur. This process is indexed by the emergence of the effect of transition probability between stem and affix (again) and/or surface frequency. The onset of LSA distance is also expected to emerge at this point, as an indicator of the system computing the semantic fit between the stem and the derived word.

The semantics-blind obligatory decomposition account also predicts that the sequence of lexical effects for pseudo-derived (opaque) word processing is equivalent to that of derived words. Naturally, form controls are not expected to demonstrate any influence of lexical properties associated with morphological processing. Thus, the same sequence of divergence points are expected to emerge as outlined for derived and pseudo-derived words, except with an absence of morphological variables.

Conversely, form-and-meaning accounts (e.g., Feldman et al., 2009; Feldman et al., 2015; Marelli & Luzzatti, 2012) predict that divergence points of semantic properties of the whole derived word would precede, or be contemporaneous with, the divergence points of markers of morphological decomposition. Specifically, form-and meaning accounts would expect to find the behavioural onset of orthographic form variables (such as orthographic neighbourhood density) to emerge early in the time course and for the emergence of these variables to be closely followed by divergence point estimates for semantic access, such as LSA distance, and other semantic variables pertaining to the whole word form (such as valence). Critically, for derived words and pseudo-derived words, this family of accounts predicts that access to meaning proceeds without intervention of morphology-specific characteristics. Thus, variables such as lemma transition probability and derivational family entropy are expected to arrive after, or concurrently with, the onset of semantic access. Again, the timeline of cognitive processing for simplex form controls is expected to follow the same sequence as derived and pseudo-derived words, with the exclusion of morphological variables. This follows from the prediction that variables associated with morphological processing are not expected to play a role in the processing of simplex words. Part of the prediction set stemming from the form-and-meaning account is that the effect of surface frequency is expected to arrive relatively early and to be indifferent to whether the word is complex or not. Therefore, for all conditions (true derived words, pseudo-derived words and form controls), on this account, one would expect to find early divergence point estimates for surface frequency.

Finally, both accounts of complex word processing would concur that a partial exposure to form would precede activation of the word's semantics in tasks where words are presented out of context. Thus, we expect the earliest influence to come from variables that are co-determined by formal properties.

We test these predictions by conducting survival analyses of (unprimed) lexical decision latencies (Studies 1-4) and eye movement fixations (Studies 5-7). Lexical decision latencies were obtained from the British Lexicon Project (BLP; Keuleers, Lacey, Rastle & Brysbaert, 2012) and the Dutch Lexicon Project (DLP; Keuleers, Diependaele, & Brysbaert, 2010). From these datasets we extracted words that conformed with three different quasi-experimental conditions: derived words, pseudo-derived words and form control words. The three conditions were selected to create a direct comparison with the conditions presented in previous masked priming lexical decision experiments (e.g., Rastle et al., 2004). Survival analysis for Dutch lexical decision was only conducted on the derived word condition. Thus, we conducted the survival analysis for the following four sets of words: (Study 1) *English derived words*, where the word consists of an English stem and suffix combination (e.g., *badness*; *bad* and -*ness*); (Study 2) *Dutch derived words*, where the word consists of a Dutch stem and suffix combination (e.g., *duiker* "diver"; *duik* and -*er*); (Study 3) *English pseudo-derived words*, cases where the word string contains a combination of a simplex word stem and an existing suffix, but where the stem does not function as a morpheme internal to the word structure (e.g., *wander*: *wand* and -*er*); and (Study 4) *English form control words*, where the whole word is simplex, contains an embedded simplex word substring, and contains a final sequence of letters that do not represent a true suffix, (e.g., *ballad*; *ball* and -*ad*). We did not perform an analysis of pseudo and form control words in Dutch because the pool of available words for these conditions in the Dutch dataset was too small for consideration.

Consideration of lexical decision latencies enables us to assess the empirical evidence for the form-then-meaning and form-and-meaning accounts that stems from (primed or unprimed) lexical decision tasks (see section entitled Evidence from masked-priming research). Our focus is on the relative order of effects and its compatibility with conflicting sets of predictions. In the same vein, we applied the survival analysis to three novel eye movement datasets (Studies 5-7) where English derived words were read in sentence context. These datasets contain eye movements to two separate experimental lists of items (Experimental list 1 shown to participants in Study 5; Experimental list 2 shown to participants in Study 6 and 7). With high temporal resolution, these data enable us to both track the relative order of effects and establish behavioral upper bounds for the time-course of morphological processing.

Lexical decision: Studies 1-4

Methods

Participants, Materials, and Procedure

Lexical decision results were retrieved from the British Lexicon Project (BLP; Keuleers et al., 2012) and the Dutch Lexicon Project (DLP; Keuleers et al., 2010), which are collections of lexical decision latencies to over 14,000 words and an equal number of non-words in English and Dutch. In both datasets, word stimuli were selected from lists of mono- and disyllabic words, representing a broad range of frequency of occurrence and phonological and morphological complexity: non-word stimuli were phonotactically valid and closely matched the syllabic and phonological structure of word stimuli. In the BLP study, 78 students and employees of Royal Holloway University of London responded to all stimuli, using their dominant hand for word responses and the non-dominant hand for non-word responses: for further details on stimuli, procedure and apparatus see Keuleers et al. (2012). In the DLP study, 39 participants responded to all stimuli using a similar procedure: for full details, see Keuleers et al. (2010).

Response variables

Lexical decision latency served as the dependent variable in all analyses for Studies 1-4. We only considered correct responses to word stimuli, which constrained the overall data pool of 2,240,940 responses to 848,108 data points for the BLP study, and from the overall pool of 1,098,942 responses to 462,244 data points for the DLP study.

Predictor variables

The key question of this paper is the relative order of effects elicited by variables related to a derived word's morphological structure (including frequency-based measures), orthography and semantics. We discuss these groups of variables in turn.

Frequency characteristics We used the 200 million-token corpus of British films and media subtitles, SUBTLEX-UK (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014) to estimate frequencies of occurrence for both whole (derived, pseudo-derived and form control) English words as well as their (pseudo)stems or embedded strings (e.g., *rainy* and *rain*; *trumpet* and *trump*; *ballot* and *ball*). The 44 million-token corpus of Dutch film and media subtitles, SUBTLEX-NL (Keuleers, Brysbaert, & New, 2010) was used for estimates of the equivalent Dutch language frequency-based measures. We refer to the frequency of the whole word as surface frequency, and the frequency of the derived/pseudo/form control stem, as stem frequency. We opted to used stem frequency instead of lemma frequency. This is because surface frequency and lemma frequency are highly correlated ($r = 0.934$), whereas surface frequency and stem frequency are weakly correlated ($r = 0.135$). Thus stem frequency is less likely to impose a confounding impact on the results of surface frequency, and vice versa. We discuss the impact of collinearity further in the supplementary materials section S1.

Morphological variables We considered three morphological variables. The first is the lemma transition probability (TPL), which is defined as the ratio of each word's surface frequency to its lemma frequency (Solomyak & Marantz, 2010). This variable measures the conditional probability of encountering the whole word given the stem, and is thus taken as an ideal variable with which to detect processing effects of morphological parsing. In addition, we computed derivational family entropy for the true derived and pseudo-derived word conditions. This was achieved by estimating lemma frequencies of words that shared the (true or pseudo-)stem with the target (e.g., *zippy* and *zipping* for *zipper*). Derivational entropy was defined as Shannon entropy calculated over the probability distribution p of the word's morphological family (obtained by dividing frequencies of family members by the cumulative family frequency): $H = -\Sigma log_2(p) * p$. Our definition of a morphological family included both derived and compound words: constraining families to just derivations did not appreciably alter results reported below. We also considered suffix productivity (SP), which was estimated as the number of word types that shared a suffix with the target word.

The metrics of morphological structure defined above (i.e. derivational entropy and suffix productivity) were quantified with the aid of the morphological parsing made available for 79,672 words in the English Lexicon Project (Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Neely, Nelson, Simpson & Treiman, 2007) and associated word frequencies from the Hyperspace Analogue to Language corpus of English (HAL; Lund and Burgess, 1996), and morphological parsing information available from the CELEX lexical database for Dutch (Baayen, Piepenbrock, & Van Rijn, 1995).

Orthographic variables To measure the orthographic neighbourhood of a word, we calculated the average Levenshtein distance (OLD20), which was defined as the mean orthographic distance from the 20 nearest orthographic neighbours. This measure was estimated for each target word in our stimulus list, using the library `vwr` (Keuleers, 2013) in the R statistical computing software program (R Core Team, 2014). A list of all unique words from the SUBTLEX-UK corpus was used as the lexicon with which to estimate orthographic neighbourhood density for a given word in English (SUBTLEX-NL was used for the same measure in Dutch). We also calculated a measure of the transition between the root and the affix in all types of words (true derived, pseudo-derived and form control words). This variable is labelled here as the bigram transition probability (TPB; see Solomyak & Marantz, 2010). TPB is defined as the frequency of the first letter of the suffix given that the preceding letter (i.e. the last letter of the stem) appears in its position relative to the end of the word. We also considered word length in characters. Finally, we considered orthographic frequency of the string that represents the target word's suffix, regardless of its morphological status within a word (e.g., the frequency of *-ly* in *rapidly* and *ply*). This measure of the frequency of the form of the suffix is referred to as form suffix productivity (FSP).

Semantic variables Following Kuperman (2013), we considered both "relational" semantic properties of derived words, i.e. ones that are defined as a relationship between meanings of the whole complex word and its embedded word, and also their "atomic" properties, i.e. ones that only require semantic access to either the stem or the whole word, and not to the relationship between them both. As a relational property, we considered a computational measure of the semantic similarity (or transparency) between the stem and the derived word. We defined semantic similarity as the cosine between distributional vectors representing two words in a multi-dimensional lexical space. To this end, we collected pairwise estimates of semantic similarity using the Latent Semantic Analysis (LSA) from the UKWAC and SUBTLEX-UK corpus of English (available at http://zipf.ugent.be/snaut-english/, Mandera, Keuleers, & Brysbaert, in press). This application uses a 300-dimensional semantic space with CBOW embeddings, and a 6-word window for calculating co-occurrence statistics. The LSA solution for Dutch was trained on SONAR-500 and subtitle corpora, and used a 200-dimensional semantic space with CBOW embeddings and a window of 10 (available at http://zipf.ugent.be/snaut-dutch/, Mandera et al., in press). A greater LSA score indicates a greater dissimilarity between the meanings of a pair of words.

One criticism of previous studies of derived words is how meanings of complex words and pseudo-complex words are discretized in order to form experimental group comparisons. For example, Baayen et al. (2011, pp. 465-466) suggested that some stimuli used in the pseudo-derived word condition of Rastle et al.'s (2004) masked-priming study may not be completely semantically opaque. For example the etymological origin of *archer* ('an individual who wields a bow') is the Latin form *arcus* ('bow'), and so the whole complex word is still structurally similar to *trucker* ('an individual who drives trucks'). Beyersmann et al. (2015) recently responded to these criticisms and created truly opaque pseudo-derived stimuli. We also imposed a more stringent set of criteria for our selection of pseudo-complex stimuli and verified that the words in this condition were indeed fully opaque, in that we excluded all *archer* examples. Furthermore, to confirm that we had indeed selected pseudo-complex words, we compared semantic transparency across each condition and found that there is no concern that pseudo-complex words carry a transparent meaning (see supplementary materials S2, where this analysis is reported). In anticipation of results, it is important to note that while we observe differences between conditions in their semantic transparency estimates, we also observe that semantic transparency, as gauged by LSA, exerts an influence on lexical decision survival rates across all word conditions. This indicates that the internal distribution of semantic transparency of each condition is sufficient to influence response patterns.

Two "atomic" semantic properties were examined: the psychological valence (positivity) of the whole word and that of its stem. Valence estimates were obtained from a set of norms to 14,000 English lemmas (Warriner, Kuperman, & Brysbaert, 2013) and a set of norms for 4,300 Dutch lemmas (Moors, De Houwer, Hermans, Wanmaker, van Schie, Van Harmelen, De Schryver, De Winne, & Brysbaert, 2013). Words were rated on a scale of 1-9 (sad to happy) by about 20 raters each. In these norming studies, words were presented in isolation, without any information about word sense, word's part of speech, or supporting context: the average of these ratings was taken as the value of the word's semantic norm. Experimental evidence has revealed that readers maintain automatic vigilance towards the emotional positivity of words (Adelman & Estes, 2013; Kuperman, Estes, Brysbaert, & Warriner, 2014). Our choice of this connotative property of word meaning was also motivated by Kuperman's (2013) observation that valence was the only semantic property of morphological constituents that exerted an effect on lexical decision latencies to compounds. That is, Kuperman (2013) found that compound words with more emotionally positive constituents (*dragon* or *fly*) facilitated RTs to the whole compound word, over and above the effect of the valence of the whole compound (i.e. the compound word *dragonfly*).

Distributional analysis

The analytical method at the forefront of this study is non-parametric distributional survival analysis, which was introduced to psycholinguistic research by Reingold, Reichle, Glaholt and Sheridan (2012) and was further refined in Reingold and Sheridan (2014). Our presentation of the method below closely follows the overview in Reingold and Sheridan (2014): we refer readers to this paper for a detailed exposition of the technique. The survival analysis method estimates the earliest point in time at which an experimental variable shows a discernible effect on a distribution of response latencies. At the focus of the survival technique is the creation of separate survival curves for conditions which are formed by levels of the experimental variable (e.g., high vs. low surface frequency, Reingold et al., 2012). A survival curve depicts, for a timepoint t, the percentage of responses (fixation durations, lexical decision latencies or other) that have a duration longer than t: see Figure 1 for survival curves of low vs. high surface frequency in Study 1 (below). In all our studies, dichotomization of all continuous variables is achieved by the median-split. The depiction of the survival curve illustrates that immediately after stimulus presentation, all responses still remain out of the full distribution of latencies as neither a button press (in lexical decision) nor the termination of a fixation by a saccade (in eye tracking) has been executed; thus for earliest timepoint, survival proportion is 100%. However, as time moves on within the distribution of responses, lexical decision button presses are executed, or eye-fixations are terminated. Thus, greater values of t are associated with a progressively higher number of trials that have been terminated by a response decision. Eventually, the survival percentage (proportion of lexical decision responses still un-executed or eye-movements that have not been terminated) decreases, until it eventually reaches 0% at t equal to the longest observed response time. If experimental conditions lead to different distributions of response times, the difference between conditions would appear as two different survival curve functions, with a faster condition (e.g., a high surface frequency condition) showing lower survival percents than a slower condition, at least at some timepoints. The distributional analysis proposed in Reingold et al. (2012) and Reingold and Sheridan (2014) identifies the earliest divergence point between two survival curves. Critically, this divergence point is indicative of the earliest temporal locus of the experimental effect elicited by the contrast between two experimental conditions.

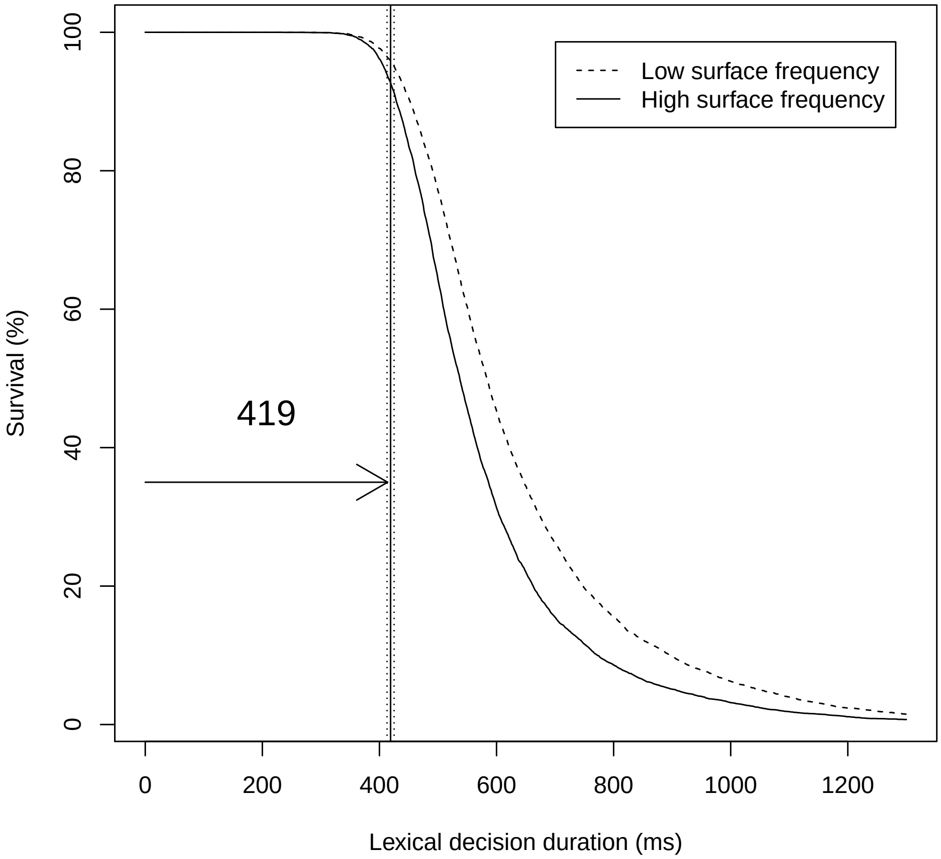

Figure 1: The divergence point estimate and its confidence interval for survival curves formed by the surface frequency effect in Study 1.

In this paper, we adopted the Confidence Interval Divergence Point Analysis (DPA) outlined by Reingold and Sheridan (2014), which produces both the divergence point estimates and confidence intervals for a comparison between experimental conditions. We computed the DPA analysis using the `RTSurvival` (Matsuki, 2016) package in the R statistical computing software program (R Core Team, 2014). The DPA procedure uses a bootstrap resampling technique (Efron & Tibshirani, 1994), which draws multiple random samples with replacement from an available pool of observations, and calculates the statistic of interest at each iteration. As implemented by Reingold and

Sheridan (2014), the DPA procedure runs 1,000 iterations of random resampling (with replacement) of response times for each participant and condition[3]. For each iteration of the bootstrap procedure, survival curves are generated for each individual participant. Next, for each 1-ms bin ranging from 1 to 1200 ms, survival percent values are averaged across participants to produce the group survival curves and the divergence point is estimated. For each iteration, the divergence point estimate is defined as the first 1-ms bin in a run of five consecutive bins in which the survival percentage in the slower (e.g., low surface frequency) condition is greater than the survival percentage in the faster condition (e.g., high surface frequency) by a pre-defined percentage threshold. Next, divergence point estimates from the 1,000 iterations are sorted from the smallest to the largest value. The 25th and 975th values of this distribution of divergence points constitute the 95% confidence interval. Finally, the median of the 1,000 divergence point values is used as the divergence point estimate for the sample.

Our studies vary in their respective sample size and statistical power. How might statistical power influence results of survival analysis? In an assessment of the DPA procedure's handling of lower statistical power (i.e. a small number of subjects and items), Reingold and Sheridan (2014) recommend 1.5% as the minimum difference threshold between survival percentages. This difference threshold avoids a false detection of a divergence point due to noise or low statistical power for as low as 12 observations per participant per condition, and Reingold and Sheridan remark that it is safe to lower this threshold in very large samples. Our datasets do not suffer from low statistical power (with about 50 to 500 observations per participant per condition). Still, to mitigate the risk of false detection even further, we stipulated a more conservative difference threshold of a 3% difference between conditions. Moreover, as can be inferred by Sheridan and Reingold (2014, Simulation 1), the accuracy of the DPA procedure for our sample sizes protects our results from a systematic bias in the relative temporal order of estimated divergence points. Finally, as an extra measure of caution, we only considered a predictor in our analyses if survival curves associated with its median-split levels showed a reliable 3% divergence in at least 700 out of 1000 iterations: we treated the remainder of predictors as unstable. In sum, these steps ensured that we uncover relatively strong effects, and further reduced the possibility of false positive detection.

To illustrate the procedure, the analysis of Experiment 1 in Figure 1 indicates 419 ms as the

[3] This aspect of the analysis requires that all participants are shown all of the experimental items. The datasets that we analyzed in this study satisfy this condition.

divergence point for survival curves of low vs. high surface frequency derived words in Study 1, with a narrow 95% confidence interval of 415 to 426 ms. This distribution of divergence points marks is a estimate of the last point in a sequence of five adjacent 1-ms bins showing a divergence of 3% or more: the median value shows that it is observed after only 5.84% of responses were terminated.

Distributional analysis applied to the timecourse of morphological processing: Our overarching interest lies in using the technique just described to detect the onset of the effect for multiple lexical and sublexical variables. For each variable in turn, we (i) generated experimental contrasts by splitting words at the median of the given predictor (e.g., whole-word psychological valence), and (ii) recorded the estimated divergence point between the two survival curves formed by responses to words in these two groups. Importantly, the non-parametric survival analysis is designed for true factorial manipulations, where all but one or two critical contrasts between stimuli (e.g., a frequency manipulation) or presentation conditions (e.g., the validity of parafoveal preview) are minimized through matching or repetition. No matching was applied to our data, which represent naturalistic distributions of variables in language use. Given the collinearity between many lexical predictors, distributions of variables other than the predictor under consideration might also vary across the two groups formed by the median split of that predictor. Furthermore, distributions of variables in surviving responses change over time. We discuss this issue in supplementary materials section S1 and show that, in our datasets, the potentially confounding effect of collinearity is not a concern.

It is also important to realize the relationship between the kind of information delivered by the survival analysis and that which is delivered by more customary analytical techniques (e.g., ANOVA or regression) which are designed to explain variance in the data. A lack of substantial divergence between survival curves estimated for factorial conditions throughout the range of latencies would translate into a small contrast between mean latencies per condition and a non-significant statistical outcome in an ANOVA. Conversely, a reliable effect in an ANOVA would imply a sufficiently large contrast between the means of response latencies to factorial conditions under comparison. Such a contrast would translate into a strong divergence between respective survival curves around the mean latency. However, it would carry no information about how early the divergence emerges and when the divergence ebbs. As we demonstrate below, the strength of a correlation between a variable and a behavioral outcome is statistically unrelated to the onset of that variable's effect.

Equally, it is possible that an effect is primarily confined to fast or slow responses, showing as an early or late divergence of survival curves, respectively. Such an effect might have a strong theoretical importance, but unless it exerts a sufficient influence on the mean of the response distribution, it would go undetected in an ANOVA (see Van Zandt, 2002 for a discussion of the drawbacks of performing ANOVAs on RTs and a detailed overview of a range of other distributional analysis techniques applied to RTs).

Furthermore, survival analysis is not designed to evaluate the contribution of a variable towards explaining variance in the data, nor can it estimate the variable's effect over and above contributions of other variables. This makes survival analysis complementary to techniques geared towards explaining variance. The former points to the onset of a variable's effect regardless of the effect size (unless the effect is too weak to elicit divergence of survival curves), while the latter carry information about other aspects of the functional relationship between that variable and either other predictors or a behavioral outcome. In this paper, we focus on the onsets of effects rather than their sizes: the latter are analyzed elsewhere (Schmidtke & Kuperman, in preparation). We do however report Spearman correlations of all predictors with dependent variables in Tables B1-7 in Appendix B.

Results and discussion

Study 1: Lexical decision of derived English words

We selected 544 derived words ending in one of the following suffixes: -*ion*, -*y*, -*ful*, -*er*, -*ness*, -*ly*, -*en*, -*less*, -*ment*, -*ize*. The derived words were relatively transparent in that their stems were free-standing lexemes and were semantically related to the meaning of the whole derived word. The full list of stimuli, along with the orthographic, morphological and semantic properties of whole words and stems, is reported fully in the online supplementary materials S3. After restricting the BLP corpus to only correct responses to the 544 derived words, a total of 19,285 response times were available. Table A1 in Appendix A reports descriptive statistics of all predictors for selected words as well as response times to the words. Table B1 of Appendix B reports pairwise Spearman correlations between predictor variables used in Study 1.

We split each predictor variable by the value of the median to create two quasi-experimental conditions. For example, splitting frequency at the median value provides two conditions, one with words that have high frequency values and one with words that have low values. The Divergence Point Analysis (DPA) was applied to survival curves for each condition in order to detect the onset of the effect that the predictor elicited: see the section Distributional analysis for the description of the procedure, and Figure 1 for the estimated temporal locus of the derived word frequency effect (419 ms, CI 95% [415 ms, 426 ms]).

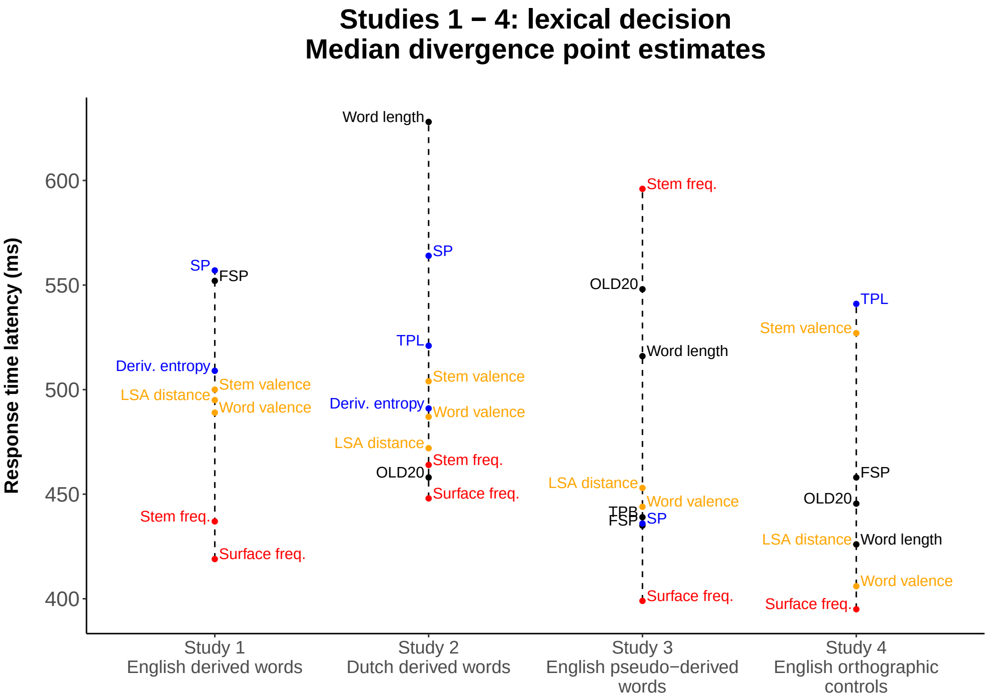

Figure 2: Plot of median divergence points for lexical decision studies: English derived words (Study 1), Dutch derived words (Study 2), English pseudo-derived words (Study 3), and form control words (Study 4). Frequencies of the whole word and stem are shown in red; orthographic predictors in black, morphological in blue and semantic in orange. Only variables with a detectable divergence point in at least 700 out of 1,000 iterations with a 3% minimal contrast are plotted. SP = suffix productivity, FSP = form suffix productivity, TPL = lemma transition probability, and TPB = bigram transition probability.

The plot in Figure 2 (refer to the timeline of Study 1 denoted on the x-axis of the plot) enables the visual inspection of median divergence points for each lexical predictor variables. For English derived words, the plot reveals that divergence point estimates of predictors occupy a broad temporal range, with median divergence points beginning at 419 ms (a point at which 5.66% of responses are terminated) and ending at 552 ms (48% of responses terminated). Surface frequency showed the earliest effect with a median divergence point of 419 ms, followed by the effect of stem frequency (437 ms)[4]. This suggests that access to the full-form precedes decomposition and morphological stem look-up. Semantic effects show their onset next, in a contiguous succession of median divergence point estimates (between 475 and 500 ms). The ordering of effects suggest that semantic access to the connotation of the whole derived word precedes that of the connotation of the stem (as indicated by effects of whole-word valence and stem valence respectively), and overlaps in time with the effect of semantic transparency. The divergence point estimates that occurred next in time were those that pertain to morphological and orthographic properties (derivational entropy, suffix productivity, and orthographic suffix probability). These effects were the only morpho-orthographic effects that passed the selection criteria and emerged late, with median divergence points between 509 ms and 557 ms. These findings suggest that orthographic cues do not play an early role in decomposing the stem and the suffix in relatively transparent "true" derived words. Moreover, access to the suffix, which should be granted as soon as the hypothesized process of early obligatory decomposition has been completed, is in fact very late, subsequent to semantic access of the whole word.

As outlined in the Distributional analysis section earlier, the quasi-experimental groups that are formed by the median split of a given predictor are not matched in our data on means or on distributions of other predictors. As is reported in supplementary materials section S1, we examined the impact of collinearity in a series of statistical tests. We found that across all of our studies (including eye movement studies) there was no cause to be concerned about the impact of collinearity.

Study 2: Lexical decision of derived words in Dutch

This study followed the same procedure as in Study 1, but concerns derived words in Dutch. We selected 74 derived words that were inflected with one of the following suffixes: -*aal*, -*aar*, -*aard*, -*ant*, -*baar*, -*dom*, -*en*, -*er*, -*heid*, -*ig*, -*ing*, -*lijk*, -*loos*, -*nis*, -*s*, -*schap*, -*sel*, -*st*, -*te*, -*vol*, -*vrij*, -*zaam*. As in the English dataset, the selection of Dutch derived words were relatively transparent in that their stems were free-standing lexemes and were semantically related to the meaning of the

[4] In our visualization of results, stem and surface are coloured in red (as opposed to blue, which marks variables associated with morphological parsing) because they signify different lexical processes for different sets of words, i.e. stem frequency in the derived word condition has a different meaning than stem frequency in form control condition. Moreover, in the morphological processing literature, stem and surface frequency are widely cited indices of access to holistic representations of parts (stem) and wholes (surface) in storage accounts (see e.g., Taft 1979; Bertram et al., 2000; Fruchter & Marantz, 2015), and so we decided to highlight them differently.

whole word (e.g., *blijheid*, meaning ‘happiness’, is composed of the stem *blij* ‘happy’ and the suffix *heid*). The full set of Dutch derived words and stems, along with their respective orthographic, morphological and semantic properties , is reported in the online supplementary materials S4. A total of 2,794 response times were available in the DLP for the set of 74 stimuli. Table A1 in Appendix A reports descriptive statistics of all predictors for selected words as well as response times to the words. Table B2 of Appendix B reports pairwise Spearman correlations between predictor variables.

As can be seen in the timeline for Study 2 in Figure 2, the results of the survival analysis revealed that surface frequency had the earliest median divergence point estimate (448 ms), which is then followed at 458 ms and 464 ms by divergence point estimates for orthographic neighbourhood density and stem frequency respectively. Convergent with the results of English derived words, this sequence of divergence point estimates suggests that access to the full-form precedes morphological stem look-up. Closely following access to the stem is the onset of a succession of three semantic effects, beginning with semantic transparency (472 ms), followed by whole-word valence (487 ms) and ending with the valence of the stem (504 ms). The earliest morphological effect, derivational entropy (491 ms), overlaps with the time window occupied by these three semantic variables. The onset of derivational entropy occurs after both whole-word and stem valence have been accessed. Following the divergence point estimate of derivational entropy are the remainder of morphological (suffix productivity and lemma transition probability), and orthographic (word length) divergence point estimates that passed the selection criterion. These morpho-orthographic effects occurred within a broad temporal range of between 521 ms and 628 ms.

Much in common with the results of English derived words in Study 1, the timeline for Study 2 in Figure 2 indicates that surface and stem frequency influenced response times at earlier timepoints than most other lexical variables. Unlike the English dataset, orthographic neighbourhood density shows an influence, and does so early, after access to the full form of the word, as indexed by surface frequency. Nevertheless, in common with English derived words, access to the suffix (suffix productivity and lemma transition probability) and the family size of the morphological family (as indexed by derivational entropy) happens after access to the meaning of the whole word and the stem respectively. Interestingly, the sequence of the block of semantic effects and the block of morphological effects were not temporally discrete, such that derivational entropy shows an onset that occurs at a timepoint that is proximate to the onset of the effect of stem valence and LSA distance. This finding may suggest that although morphological decomposition does not precede semantics temporally, the processes of decomposition and semantic access may be initiated concurrently.

In sum, these findings demonstrate, in conformity with the results from English language lexical decision dataset, that morphological parsing takes place only once access to the whole word form has been activated. Indeed, it is logical to assume that the orthographic identification of a word string (either via its surface form or its stem) is a necessary initial step toward accessing a word's meaning. Crucially, the results so far suggest (for Dutch and English derived words) that access to a complex word's meaning does not wait until after obligatory decomposition of the word string has taken place. Instead, morphological decomposition of the complex word appears to occur, at its very earliest, simultaneously with access to the whole-word meaning.

Study 3: Lexical decision of pseudo-derived words

The study to which we now turn focuses on semantically opaque (pseudo-)derived words, e.g., words containing an existing suffix and a semantically unrelated stem (*broth* and *-er* in *brother*). Identifying divergence point estimates for this variable was motivated by the presence of the same condition in masked priming studies (see e.g., Rastle, Davis, & New, 2004). The original reasoning behind the inclusion of this condition in these experiments was to test the hypothesis that morphological decomposition operates independently of semantic information. For example, according to Rastle et al. (2004), "Any stimulus bearing a morphological surface structure ... would be decomposed, irrespective of its semantic transparency or etymological characterization" (pp. 1091).

We proceeded to test this hypothesis by identifying 102 pseudo-derived English words that had available behavioural latencies in the BLP. We selected words which had a range of sublexical and lexical predictors for the stem, suffix and the entire word. These lexical characteristics were the same as were collected for the English and Dutch derived words. The full list of pseudo-derived word stimuli, along with the properties of whole-words and stems, is reported in the online supplementary materials S5. As stated earlier, we did not perform a comparable analysis in Dutch because the pool of pseudo-derived words in the Dutch dataset was too small for consideration. After the trimming procedure described above, these words yielded a pool of 4,976 response times from the BLP. Table A1 in Appendix A reports descriptive statistics of all predictors for selected words as well as response times to the words. Table B3 reports pairwise Spearman correlations between predictor variables.

Firstly, as is illustrated in Figure 2 (Study 3 timeline), the range of median divergence points is larger than in all other conditions. In addition, the onset of the earliest variable to influence response times takes place before that of both of the true derived word conditions. Once again, surface frequency emerges as the earliest variable to exert an influence, at 399 ms. The very last variable to demonstrate a divergence in lexical decision times is the frequency of the stem, some 200 ms later, at 596 ms. The late influence of stem frequency may be indicative of the reduced role of the meaning of the stem for the process of decomposition, and the non-obligatory nature of early decomposition. This finding will be discussed further in the General discussion.

Following the onset of surface frequency, the next group of variables to demonstrate an impact on lexical decision pertain to either the orthographic or morphological characteristics of the word: form suffix productivity (435 ms), the frequency of the suffix's status as a true morphological string (suffix productivity, 436 ms) and the probability of encountering the first letter of the affix given the last letter of the root (bigram transition probability, 439 ms). Crucially, the temporally compact onsets of these variables coincide with the emergence of the effect of whole-word valence (444 ms) and semantic similarity (453 ms). The simultaneous accompaniment of morphological and semantic effects is further visualized in the vincentile plot (Ratcliff, 1979; Vincent, 1912) of Figure S9.3 in supplementary materials S9. Unlike Figure 2, which plots the median divergence point estimate for each variable, Figure S9.3 depicts, for each variable, the full range of divergence point estimates within the 95% confidence interval (for full details on plotting technique, please consult explanation in supplementary materials S9). The plot shows that, even at the earliest possible divergence, morpho-orthographic and semantic effects show considerable temporal overlap. Moreover, this overlap remains stable across the full range of divergence points. To sum up, the virtually contemporaneous onsets of lexical characteristics linked to morphological parsing, and those related to the semantics of the whole word, suggest that morphological decomposition of pseudo-derived words is not indifferent to the role of semantics.

Study 4: Lexical decision of form control words

To recap, previous masked priming studies have detected no priming differences between true derived and pseudo-derived (opaque) words, while the magnitude of priming in these conditions was greater than those for form control words (Rastle & Davis, 2008). Although this study does not follow the same design and methodology as a masked priming study, we opted to include a baseline condition with which to compare divergence point estimates with those of true and pseudo-derived words. To this end, this study considered a set of 158 words that are not morphologically complex (i.e. do not contain a true affix i the English language), but begin with a substring that is a simplex word. For example, the simplex word *chapel* contains the substring *chap*. Two constraints were applied during the selection of these words. The first was that the whole-word string was, at maximum, three letters longer than the embedded word. The second was that those final letters (e.g., -*el* in *chapel*) were not true English suffixes. There were also two variables from the previous three studies that could not be estimated for the current set of words, by virtue of these words not being morphologically complex. These variables were derivational entropy and suffix productivity.

A total of 5,350 response times were available in the BLP for the selected stimuli. Table A1 in Appendix A reports descriptive statistics of all predictors for selected words as well as response times to the words. Table B4 of Appendix B reports pairwise Spearman correlations between predictor variables. The full list of stimuli, along with the properties of whole-words and stems, is reported in the online supplementary materials S6. We did not perform the same analysis in Dutch because the pool of simplex form control words in Dutch data was too small for consideration.

Comparable with the pseudo-derived word condition, Figure 2 (Study 4 timeline) reveals that the first divergence point is earlier than for both true derived word conditions. Moreover, as with all conditions, surface frequency is the first variable to exhibit an influence over lexical decision latencies, with a median divergence point estimate of 395 ms. Surface frequency is then followed by whole-word valence at 409 ms, and then LSA similarity at 426 ms. As the vincentile plot in Figure S9.4 of supplementary materials section S9 shows, whole-word valence and word length share often overlapping onsets throughout the full distribution of divergence point estimates. That is, word length has a median divergence point estimate that is identical (426 ms) to the divergence point estimate of whole word valence. However, Figure S9.4 also shows that the earliest divergence point estimate for word length (399 ms) precedes that of word valence (411 ms). Following this pairing of variables are the median divergence point estimates of two orthographic variables: orthographic density (445 ms) and form suffix productivity (458 ms). Finally, the onsets of the effect of stem valence and lemma transition probability (defined as the ratio of the word's surface frequency to the frequency of its embedded string) arrive last, with median divergence points of 527 ms and 541 ms respectively.

Overall, it is apparent in the summary plot in Figure 2 (Study 4 timeline) that only one variable associated with morphological parsing passed the criteria that we set for the survival analysis. This result is expected as all but one of the morphological variables were not available for this word group from the outset. When the effect of lemma transition probability (the only 'morphological variable') does arrive, it does so late. This is also not surprising; because the substring is not morphologically related to the whole word, the computation of the frequency of the substring relative to the frequency of the whole word is not a high priority for the language system.

Moreover, it again appears that semantic access begins to exert an influence on response time latencies relatively early in the time-course, and once again suggests that once frequency of the whole form has been activated, language processing immediately becomes sensitive to what the letter string may signify. This reasoning is supported by the early onset of the effect of the valence of the word, which is then followed by LSA distance. In addition, this finding suggests that even for simplex words, the meaning of an embedded substring is evaluated during the recognition of the whole word. This converges with the semantic effects of word strings embedded within larger simplex words reported by Bowers, Davis, and Hanley (2005). Lastly, it is intriguing that, in convergence with the pseudo-derived condition, access to the semantics of the word is not mediated by the frequency of the stem. Whereas Study 3 eventually suggested a late effect of stem frequency, the influence of this variable did not pass the criteria that we set for the survival analysis here. This absence may be indicative of a difference in decomposition processes across conditions.

To sum up, the key results of Studies 1-4 is that surface frequency demonstrates the earliest detectable effect in survival analysis, and is constant across all of these studies. Moreover, most semantic variables show early divergence points that are either simultaneous with or precede those of orthographic variables. Finally, a morpho-orthographic variable of lemma transition probability shows the latest influence on lexical decision latencies out of all lexical variables.

Even though eye tracking data, as compared with lexical decision, places stricter constraints on the temporal upper-bound of neural activity (see below), it is interesting how the time windows proposed for semantic effects in the brain imaging studies of derived words reconcile with the lexical decision results presented in Studies 1 to 4. The average latency across conditions in the lexical decision data considered here is on the order of 550 ms. Yet, for factors like word frequency and semantic transparency to affect lexical decision responses, one expects the influence of these factors to be sufficiently strong before the response is initiated. The average latency of a hand movement in the lexical decision task is estimated at 130 ms (Balota & Abrams, 1995). Thus, one expects to see signatures of word frequency and semantic transparency in neural activity at, or prior to, the upper temporal bound of 420 ms (550 ms - 130 ms). This estimate is consistent with the earliest effects observed in our survival analysis (surface frequency effect at 400-420 ms post-onset). It also compatible with the semantic onset of 260 ms proposed some in neuro-imaging studies (Lavric, Elchlepp, & Rastle, 2012 and Zweig & Pylkkänen, 2009). Yet it can hardly be reconciled with the time windows of 350-500 ms and 400-600 ms proposed in other neurophysiological research reviewed above.

In the next set of studies we turn to the eye tracking methodology, which enables one to approximate the upper bound of neural activity associated with derived word processing with a more fine-grained temporal resolution. We demonstrate an even more drastic discrepancy between behavioral and neural signatures of morphological processing.

Eye tracking: Studies 5-7

The eye movement component of this paper is composed of three separate eye tracking studies. Two experimental lists of English derived words (Experimental lists 1 and 2) were generated and embedded in sentence contexts. Experimental list 1 was shown to one group of participants (Study 5) and Experimental list 2 was shown to two different population samples (Study 6 and 7). Recruitment of participants in Studies 5 and 6 relied on convenience sampling and both sets of participants were sampled from the same university population. Participants that were tested in Study 7 were sampled from the local community in Hamilton, Ontario, Canada. These participants were sampled in order to represent adult non-college bound readers. They were recruited as part of a larger study which aims to assess hypotheses pertaining to individual differences in reading ability. These studies were approved by the local McMaster Research Ethics Board (McMaster University). Below we detail the methods for each study.

Methods

Participants

Study 5 34 undergraduate students from McMaster University (25 female; 9 male) within an age range of 18-37 (M = 19.91, SD = 3.33) completed the eye tracking study for course credit. All participants were native speakers of English. All participants had normal or corrected-to-normal vision, and did not report a diagnosed reading or learning disability.

Study 6 38 undergraduate students from McMaster University (27 female; 11 male) within an age range of 18-30 (M = 20.82, SD = 2.7) completed the eye tracking study for course credit. All participants were native speakers of English. All participants had normal or corrected-to-normal vision, and did not report a diagnosed reading or learning disability.

Study 7 45 participants (23 female; 22 male) were recruited in Hamilton, ON, Canada, within an age range of 18-31 (M = 23.24, SD = 4.21). Participants were paid $15-20 CAD/hour and were recruited from the local community in a number of ways, including presentations at local colleges; advertisements placed on local community sections of the online classified advertising services (Craigslist, Indeed and Kijiji); posters/flyers placed on adult school and community college campuses, public transportation hubs, and from referrals from past and current study participants. All participants were non-college bound individuals (formal level of education did not exceed the equivalent of high school level). All were native speakers of English, had normal or corrected-to-normal vision and none had a diagnosed reading or learning disability.

Materials

Experimental lists 1 and 2 each consisted of 200 unique English derived words (e.g., *lockable*). The words were each embedded within a single sentence context (e.g., *The small clockwork mechanism had lockable hinges.*). For Experimental list 1, we selected derived words with the following suffixes: -*ion*, -*er*, -*ment* and -*ness*. For Experimental list 2, we selected derived words with the following suffixes: -*able*, -*ful*, -*ity* and -*ive*. For both lists, the sentence context preceding each derived word was neutral and each derived word did not occupy the first or last position of each sentence. All sentences were limited to 90 characters in length and did not exceed one line on the computer screen.

Not all eye movements to stimuli in each list could be used in the survival analyses. This is because not all the independent lexical variables were available for all of the words in each experimental list (see Predictor variables subsection in the Methods of Studies 1-4 for a detailed list of these variables). For Experimental list 1, 101 items of the total list of 200 word items were attested with a complete set of lexical variables. For Experimental list 2, a full set of lexical variables were available for 145 words out of the entire 200 word list[5]. The full list of materials from Experimental lists 1 and 2 for which there was a complete selection of lexical variables is reported in the online supplementary materials S7 and S8, along with the lexical properties of whole-words and stems.

Apparatus and procedure

All eye tracking studies followed the same procedure. The sentences were displayed on a 17-inch monitor with a resolution of 1600 x 1200 pixels, and a refresh rate of 60 Hz. Eye movements during sentence reading were recorded with an Eyelink 1000 desk-mounted eye tracker (SR Research Ltd., Kanata, Ontario, Canada). The eye tracker is an infrared video-based tracking system combined with hyperacuity image processing. The data were collected at a 1000 Hz sampling rate from the participants' dominant eye, or the right eye if the dominant eye was not known. Sentences were presented one at a time in Courier New, a monospace font, size 20, in black on a white background, and occupied exactly one line on the screen. Each character subtended 0.36° of visual angle. A three-point horizontal calibration of the eye tracker and a three-point horizontal accuracy test were performed before the beginning of each experiment, and after any breaks. A chin support and forehead rest was used to stabilize participants' gross head movements.

Each experiment began with a practice block, consisting of ten sentences, in order to familiarize participants with the experiment. Participants then silently read sentences containing the target derived words in their sentence context. Participants were instructed to press a button when they

[5] For the sake of completeness, we report that the list of derived English words only partially overlapped with the words used in the lexical decision experiment reported in Study 1. That is, 32 items from Experimental list 1 and 22 items from Experimental list 2 were also present in Study 1's stimuli list.

had finished reading the sentence, and the sentences remained on the screen until the button was pressed. For each experiment, participants read the 200 target sentences in randomized order. Each sentence trial was preceded by a drift correction, which used a fixation point positioned 20 pixels to the left of the beginning of the sentence, in order to ensure accurate recording of eye movements. Sentences were presented 100 pixels away from the left edge of the screen, and in the middle of the vertical dimension of the screen. Comprehension questions followed 20% of target sentences. Participants were presented with the sentences and were asked to respond whether they were true or false. Participants pressed the "a" key if the sentence was true and the " ' " (single quote) key if it was false. 50% of the correct answers were true, and 50% were false. The proportion of correct responses across all three studies was high $> 93\%$.

Response variables

The dependent variable for the eye movement analyses was the first fixation duration. This dependent measure is based on all trials in which there were no more than two fixations on the critical derived word in the first pass of reading, i.e. the vast majority of trials.

Predictor variables and statistical analyses

The same lexical measures were collected and computed as in the derived word condition for the lexical decision study (Study 1), see the Predictor variables section in the Methods for Studies 1-4 for a full description. Table A2 in Appendix A reports descriptive statistics of all predictors and for first fixation duration times for Studies 5-7. Tables B5-7 in Appendix B report pairwise Spearman correlations between predictor variables and first fixation duration times for Studies 5-7.

The same survival analysis procedure used in Studies 1-4 was applied to Studies 5-7: see the method in Distributional analysis above. The only exception was that we relaxed our criteria for what counts as a stable predictor: we considered predictors that showed reliable divergence points in at least 500 out of 1000 iterations (as opposed to the minimum of 700 in lexical decision studies). The adjustment is due to relative weakness of effects on eye-movement durations. As in the lexical decision studies, we checked for a confounding impact of collinearity on divergence point estimates and did not observe any: see supplementary materials section S1.

Results and discussion

The same data clean-up procedure was applied to each eye movement data set. We removed trials for which the eye tracking signal was lost, the target word was skipped, was fixated on for more than 6 times, was fixated on for the first time after gaze proceeded past the target word, was fixated on for less the 80 ms or was fixated on for more than 1000 ms.

For eye movement Study 5, the initial raw data set consisted of 3,354 trials. The application of data cleaning procedures led to a loss of 894 (36.3%) trials from the initial raw data set. The resulting final data set comprised of 2,460 valid trials. For eye movement Study 6, we began with a raw data set of 5,247 trials. The data cleaning steps resulted in a loss of 1,074 (25.7%) data points, which produced a final data set of 4,173 valid trials. Finally, for eye movement Study 7, we started out with a raw data set which contained 6,253 trials. The application of data cleaning steps reduced this data set to 4,186 valid trials (1,061 trials - 25.3% of data points lost). After clean-up the proportion of trials for which a second fixation was made was 46.71% in Study 5, 44.02% in Study 6 and 53.51% in Study 7. For presentational purposes, we collapse together the reporting of results for each study.

The plot in Figure 3 enables the visual inspection of median divergence points for all lexical predictor variables across Studies 5 to 7. The eye movement results show that onsets of lexical variables occur some 300 ms earlier than onsets for the same variables in lexical decision. The earliest median divergence point estimate is at 141 ms in Study 5 (a point at which 7.03% of eye fixations have terminated, i.e. a saccade was executed and the eye either refixated on the word or regressed out of the word) and the latest divergence point estimate is also found in Study 5 at 255 ms (69.67% of eye-fixations terminated). Although dramatically shifted to an earlier absolute timeframe, the range of median divergence point estimates from the earliest lexical variable to the latest lexical variable (114 ms) is consistent with that of Study 1 (133 ms). This finding suggests that the time window during which the earliest discernible influence of all lexical variables are detected is comparable across eye-movement and lexical decision experiments.

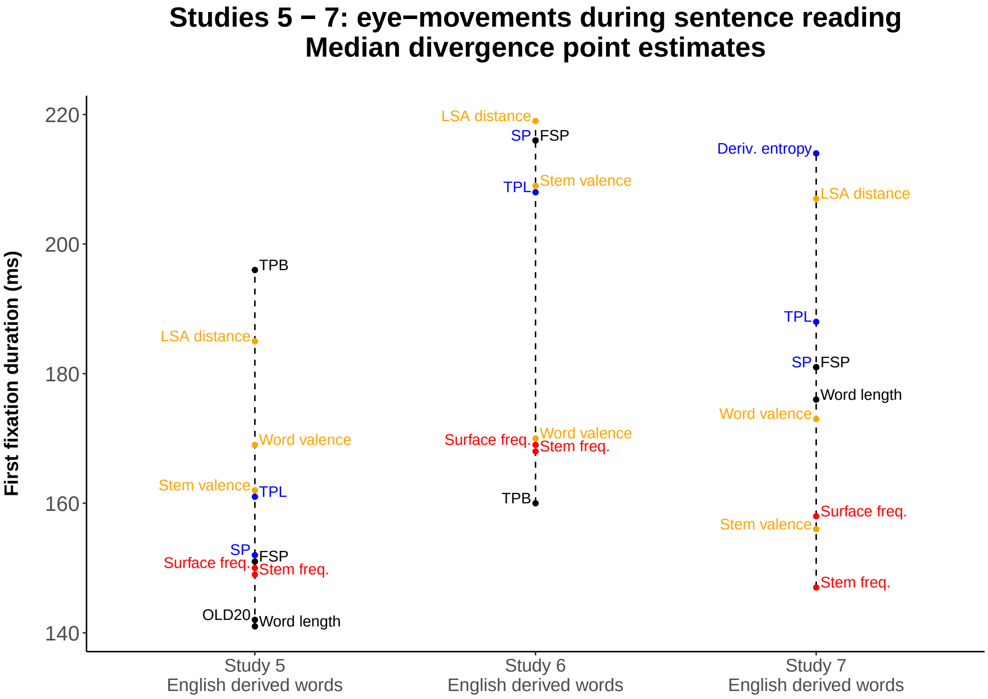

Figure 3: Plot of median divergence points for first fixation durations to derived words in English which were presented in sentence context. Words from Experimental list 1 were included in Study 5, and words from Experimental list 2 were included in Study 6 and 7. Frequencies of the whole word and stem are shown in red; orthographic predictors in black, morphological in blue and semantic in orange. Only variables with a detectable divergence point in at least 500 out of 1,000 iterations with a 3% minimal contrast are plotted. SP = suffix productivity, FSP = form suffix productivity, TPL = lemma transition probability, and TPB = bigram transition probability.

Firstly, across all studies the results of the survival analysis indicate an early onset of both stem frequency and whole-word frequency. The onset of stem frequency occurs at 149 ms, 168 ms, and 147 ms for Studies 5 to 7 respectively. These median divergence point estimates accompany those of surface frequency, which arrive at 150 ms, 169 ms and 158 ms for Studies 5 to 7 respectively. For Studies 5 and 6, the onsets of variables related to orthographic structure of derived words precede or accompany the divergence points of stem and surface frequency: word length (Study 5; 141 ms), orthographic neighbourhood density (Study 5; 142 ms), form suffix productivity (Study 5; 151 ms) and bigram transition probability (Study 6; 160 ms). Collectively these results lead to two observations: (i) processing of the whole word, the stem and the suffix begins simultaneously, in line with the dual- and multiple-route accounts and contrary to obligatory decomposition accounts; and (ii) effects of whole-word and stem frequency have a substantial orthographic component, because they emerge at the time when other pure orthographic influences emerge too. As in the results from lexical decision, the observed early effect of surface frequency is not compatible with the late onset of its effect in accounts proposed by Solomyak and Marantz (2010, see quotation 1. above), nor in Fruchter and Marantz (2015) or Taft (2004). Therefore, the results of these studies uncover a chronology of effects in which access of the orthographic structure of the word and access of the surface form (as indexed by surface frequency) is not mediated by morphological parsing of the stem and affix (which would be indexed by morphology-specific variables). In other words, we do not find that the onsets of variables defined as indices of morphological parsing come between access of purely orthographic properties (variables coded in black in Figure 3) and surface frequency. This time-line is inconsistent with the chronology of decompositional processes put forward by, among others, Marantz and Solomyak (2010) and Fruchter and Marantz (2015).

Secondly, the relative ordering of morphological and semantic effects are largely consistent with the above lexical decision results. In Study 6 and 7, the respective divergence point estimates of valence of the whole word arrives at an early absolute timepoint (170 ms and 173 ms), and precedes those of lemma transition probability (208 ms and 188 ms) and suffix productivity (216 ms and 181 ms). Furthermore, derivational entropy, another variable which reflects morphological parsing, produced a reliable divergence point estimate for one study (Study 7; 255 ms). As can be seen in Figure 3, this was the latest divergence point estimate across all eye movement studies. In Study 5 however, the morphological structure of derived words did demonstrate relatively early divergence points. Both suffix productivity (152 ms) and lemma transition probability (161 ms) showed their influence slightly before or at the same time as effects of valence of the stem (162 ms) and the whole word (169 ms). We conclude that – as in Studies 1-4 – access to morphological properties of the word and its morphemes does not reliably precede access to their semantic properties.

Interestingly, in line with the predictions of the semantics-blind obligatory decomposition account (see Fruchter & Marantz, 2015), we observe late effects of LSA distance (semantic transparency) relative to variables pertaining to parsing of the stem and affix (e.g., lemma transition probability and suffix probability). We argue that the late effect of LSA distance may be an indicator of a post-access attempt to compute the semantic similarity between the stem and whole-word meanings.

To sum up, the results of the eye movement studies bring forth three key findings. First, as with lexical decision, surface frequency has a relatively early (150-170 ms) influence on first fixation durations during English derived word recognition. This onset coincides in time with effects coming from formal properties of the stem and suffix, suggesting simultaneous access to morphemes and the whole word. Second, in Study 6 and 7 the valence of the stem and whole-word also show divergence points that preceded or were contemporaneous with those of lexical variables which are taken to index morphological parsing. That is, semantic effects were found earlier or as early as pure morphological effects. Finally, the absolute time-points of the onset of these effects in each eye movement study do not agree with many of those reported in the neuro-imaging literature. It is possible to conclude that, collectively, the eye-movement results presented in Studies 5 to 7 do not substantiate a temporal flow of information that would be expected under the form-then-meaning account of morphological decomposition. We elaborate on these findings further in the General discussion.

General discussion

The aim of this study is to provide scrutiny of the current accounts of the time-course of morphological processing in visual word recognition tasks. An influential form-then-meaning account of morphological processing proposes that the recognition of any word ending in a decomposable suffix follows a strict staged sequence of automatic cognitive processes, and does so whether that word is truly derived (e.g., *dreamer*), or is pseudo-derived (e.g., *corner*). The sequence entails fast and obligatory decomposition of the derived word into its morphemes (stem *dream* and affix *-er*), lexical access to the stem and licensing of its combinability with the suffix, and late recombination of the meanings into a unified semantic representation (see Introduction for references). This hypothesized sequence predicts that specific lexical variables will show influence at certain stages of the timeline of complex word recognition. These diagnostic variables, and the expected relative order of their influence, are outlined both in the behavioral and neurophysiological literature (cf. quote 1. from Solomyak & Marantz, 2010, and, among others, Beyersmann, Ziegler, Castles, Colheart, Kezilas, & Grainger, 2015; Fruchter & Marantz, 2015; Marslen-Wilson, Bozic, & Randall, 2008; Rastle, Davis, & New, 2004; Taft, 1979, 2004). Furthermore, reports of brain activity also associate processing stages with absolute estimates of time windows in which respective variables exert their influence.

We reviewed this proposal in light of existing behavioral evidence on complex word recognition (see Introduction) and harnessed a novel distributional approach to the analysis of reaction time data. The analysis that we adopted is specifically geared towards our aim of shedding light both on the relative order of behavioral stages and their absolute loci in the time-course of processing. Specifically, we applied the survival analysis technique to lexical decision latencies to a large selection of true derived words in English and Dutch, as well as English pseudo-derived and form control words (Studies 1-4). We also applied the same analytical method to three eye movement datasets containing English derived words presented in sentence context (Studies 5-7). The survival analysis produces estimates of divergence points between survival curves associated with levels of variables of interest (e.g. high vs. low surface frequency, see Figure 1). These divergence points indicate onsets of predictors' effects in the behavioral record, i.e. timepoints at which the influence of a predictor on recognition behavior is appreciable. As such, divergence points serve as upper temporal bounds for the expected absolute timing of respective effects in brain activity: since neural activation is a cause of behavior, it has to precede behavioral manifestations. Thus, the relative order of divergence points serves as both a test of the empirical evidence that the form-then-meaning account garnered in the behavioral literature and a constraint on the expected temporal sequence of processing stages in the brain. In what follows, we consider the implications of the lexical decision and eye tracking findings for the absolute and relative characteristics of the time-course of morphological processing. While we acknowledge the existence of a large body of the masked priming literature on the topic of morphological processing (see Introduction), our discussion below concentrates on comparing our findings to the paradigms that afford a direct high-resolution measurement of temporal processes, i.e., the brain imaging findings.

The absolute time-course of morphological processing

Both our review of the eye tracking literature (Introduction) and our own findings (Studies 5-7) run counter to the proposed time-course that is based on neurophysiological data. The analysis of average fixation latencies in complex words where frequency-based and semantic effects (e.g., surface frequency or semantic transparency) were observed pointed to the time window of 170-250 ms as a temporal locus of these effects and an upper temporal bound for corresponding effects on neural activity. In our review of the literature, we found that this estimate held true both for complex words presented in isolation and in context. The results of survival analysis presented here shifted this time window to an even earlier point. This time window, from 140 to 250 ms, contained onsets of all effects of behavioral influence, formal and semantic. These findings are clearly at odds with the neural signatures of semantic effects on derived word processing, as reported in brain imaging research, i.e. 260 ms post-onset (Lavric, Elchlepp, & Rastle, 2012 and Zweig & Pylkkänen, 2009), or 350 or 400 ms post-onset (Beretta et al., 2005; Lewis et al., 2011; Simon et al., 2012; Solomyak & Marantz, 2009; 2010). It is however fully in line with the timeline of simplex and complex word processing in numerous EEG and MEG studies (Assadollahi & Pulvermüller, 2003; Hauk et al., 2006; Jared et al., 2016; Pulvermüller, 2002; Reichle, et al., 2011; Sereno et al., 1998). (We also discuss a similar discrepancy between the survival analysis of lexical decision data and some neurophysiological reports at the introduction of the eye-movement section.)

To reconcile the discrepancy between literatures, one either has to conclude that (a) the results of over a dozen eye tracking studies systematically produce spurious early semantic and frequency effects on initial eye fixations on words in context and isolation, or that (b) the timing of the onsets of semantics and frequency effects reported in some of the neurophysiological studies do stand, thus breaking a relationship of cause-and-effect between neurophysiological activity and its associated behavioral response, or that (c) temporal estimates of brain activation related to derived word semantics are erroneous. We submit that there is no evidence for (a) and no logical basis for (b). We conclude that a likely source for discrepancy lies in (c). In our case, (c) is expressed in the implausible temporal estimates originating from analyses of MEG and EEG data. We remind the reader that conclusion (c) might also be expressed as the absence of detected neural counterparts of behavioral effects. Recent studies co-registering the eye tracking and EEG data (Dimigen et al., 2011; Kretzschmar et al., 2015) demonstrate both kinds of error. They observed that behavioral signatures of benchmark effects of processing effort (predictability or frequency) either preceded neural signatures of those effects or did not have corresponding neural signatures at all (for further discussion of this apparent paradox see Dambacher & Kliegl, 2007; Dimigen et al., 2011; Kliegl, Dambacher, Dimigen, Jacobs, & Sommer, 2012; Sereno & Rayner, 2003; Sereno et al., 1998). For an in-detail discussion of causal relations between research of brain and behavior, see Krakauer, Ghazanfar, Gomez-Marin, Maclver, & Poeppel, 2017).

We do not pass judgment on what aspect of experimental design, apparatus, data collection or analysis of brain imaging data requires revision. However, we make a contribution to this field of research by estimating upper temporal bounds for when the stages of morphological processing ought to have a signature in the brain activity record. We believe co-registration of behavioral and neutral activity within-participants to be an optimal methodological solution for the future effort of developing credible temporal estimates from brain imaging data.

Relative order of effects on morphological processing

Under the form-then-meaning account, affix-specific variables, such as affix frequency, are predicted to exert influence first in the time-course of derived word recognition. Next in line are variables related to parsing of the stem and affix, such as transition probability between the stem and affix. Thus, the initial stages of word recognition are expected to be dominated by variables indicating access to orthographic form and morphological structure. The next stage is arguably devoted to lexical access of the stem, and is predicted to be contemporaneous with the onset of the effect of lemma frequency, stem frequency or derivational family entropy (Fruchter & Marantz, 2015). Following morphological parsing and stem look-up, recombination of the stem and affix begins to take place. According to Solomyak and Marantz (2010), this stage may be indexed by the emergence of an additional effect of transition probability between stem and affix, LSA distance (as an indicator of the system computing the semantic fit between the stem and the derived word) and ultimately surface frequency. Crucially, this final recombination stage is where – under this account – the semantics of the word is accessed and word recognition has been achieved.

No diagnostic feature of fast, obligatory and semantically blind decomposition found consistently reliable support in our survival analyses of lexical decision or eye tracking data: i.e. access to morphological structure prior to access of whole-word semantics, and late involvement of surface frequency. First, contrary to the predictions above, the temporal emergence of semantic effects preceded those of morphological parsing during the processing of English and Dutch derived words in the unprimed lexical decision task (Studies 1-2) and also in the silent reading of English derived words (with a partial exception of Study 5 where these effects were roughly simultaneous). Specifically, effects of stem and whole-word valence were reliably observed prior to (or in Study 5 at the same time as) effects of morphological structure, e.g., derivational entropy or lemma transition probability. Since valence is an atomic semantic property, i.e. a property defined for a lexical unit independently of whether it is embedded or embeds other units, this is direct evidence that access to semantics of derived words and their stems is not delayed until formal morphological decomposition is over.

Interestingly, in Study 3, processing of English pseudo-derived words also revealed onsets of semantic effects that were at least contemporaneous with onsets of lexical characteristics linked to morphological parsing and the statistical properties of letter strings (see also the divergence point time-course in Study 1, Study 2 and Study 3 in Figure 2, and also in Study 6 and 7 in Figure 3). This finding suggests that at the stage subsequent to accessing form, the lexical properties relevant to processing may also be semantic and not just morphological in nature.

Second, across all studies, surface frequency showed one of the earliest divergence points in survival curves and – except Study 7 – preceded or was concurrent with the onset of the stem frequency effect. According to a semantics-blind account of obligatory decomposition, the initial lexical search stage involves access to the stem of the complex word (e.g., *dream*), and a later stage involves access to the whole-word unit (e.g., *dreamer*). Thus, under this account, constituent frequency effects are attributed to the early stages of morphological decomposition, whereas surface frequency effects are attributed to the subsequent recombination and semantic access (Solomyak & Marantz, 2010; Taft, 2004; see also Baayen, 2014). In the current lexical decision studies, contrary to the account's predictions, divergence point estimates of surface frequency arrive at the earliest time out of all variables, and naturally, preceded divergence point estimates of stem frequency in Dutch derived words, and in English derived and pseudo-derived words (stem frequency does not influence reaction times in the form control condition). In eye-movements too, surface and stem frequency demonstrated very early and roughly contemporaneous divergence points, indicative of simultaneous access to words and morphemes, contrary to premises of the single-route obligatory decomposition approach but compatible with a family of other accounts. As argued in Kuperman (2013), surface frequency is both a semantic and a formal property of a word. We believe its observed early engagement in word processing points to the formal aspects of what surface frequency indicates, e.g., one's familiarity with the pattern of printed symbols constituting the word through repeated exposure, which enables orthographic processing even in nonhuman primates (Grainger, Dufau, Montant, Ziegler, & Fagot, 2012). We find additional support for this notion from the observation that in eye-tracking studies the divergence points of either surface or stem frequency coincided with those of purely formal word properties, e.g., orthographic neighbourhood density, word length, or bigram transition probability.

Another crucial finding is that the overall pattern of results for pseudo-derived words are largely convergent with the results of the true derived word conditions. For pseudo-derived words, the results indicate a rapid response to the form of the whole word (as indexed by surface frequency), which is then followed by concurrent access to semantics (whole-word valence and LSA distance) and to morpho-orthographic variables (form suffix productivity and suffix productivity). Interestingly, unlike true derived words, the results suggest that frequency of the stem is not registered until very late and access to the family of the stem (derivational family entropy) is not attempted (see Figure 2). It is possible that these effects are evidence of a weak influence of the stem meaning during the processing of the whole opaque word (i.e. accessing the meaning of *trump* in *trumpet*). This speculation is in line with the findings that show that semantic opacity modulates complex word processing (see e.g., Juhasz, 2007; Feldman et al., 2009; Marelli & Luzzatti, 2012). Moreover, a comparison across timelines for true derived and pseudo-derived words (Studies 1-3 and 5-7), and simplex words (Study 4) suggest that the processing of complex and simplex words does not differ as radically as the obligatory decomposition account suggests (also, see above for similarities in the temporal placement of semantic effects in true and pseudo-derived words). Regardless of real or apparent morphological structure, words in all conditions show very early effects of surface frequency (possibly reflecting the ease of recognizing the given string of characters) and whole-word valence (reflecting automatic vigilance for dangerous and useful objects, see below). This implies that morphological structure may enable one to use additional processing routes (such as decomposition into morphemes, or separate access to and subsequent recombination of meanings). Yet, even if this were true, we find little evidence that (apparent or true) word structure elicits an entirely different strategy of visual recognition, with its own attentional and cognitive mechanisms posited by some of theoretical accounts.

In sum, the attested relative order of frequency-based, morpho-orthographic and semantic effects in the time-course of derived word processing runs counter to central predictions of the form-then-meaning account and the masked priming and unprimed lexical decision studies that form its basis (see Introduction). We do not find support for the fast, obligatory and semantically blind nature of decomposition of truly or potentially complex words into their morphemes, followed by access to the stem meaning and its recombination with its suffix. Our results however, are in line with any account that argues for simultaneous engagement of all sources of information – formal or meaning-related – in the task of identifying a word in print. For instance, our findings are compatible with the dual- or multiple-route theories of morphological processing, which argue for a parallel and interactive use of properties associated with whole words and their morphemes (Burani & Caramazza, 1987; Chialant & Caramazza, 1995; Lehtonen, Vorobyev, Hugdahl, Tuokkola & Laine, 2006; Kuperman et al., 2009; Schreuder & Baayen, 1995). Our findings also agree with the proposal of Grainger and Ziegler (2011) that access to complex words is possible through a fast track, i.e. direct access to orthography and then meaning of the whole word, or through a slower process of morpho-orthographic decomposition. Finally, our results are also compatible with the recent Naive Discriminative Learning model (NDL; Baayen et al., 2011), which views morphological processing as dynamic learning of mappings between formal cues (including orthography and phonology of the whole word, its embedded morphemes and n-grams) and meanings. The NDL model moves away from a storage based metaphor of accessing a stored meaning, to a view of word processing in which strengths between formal cues and potential meanings are in constant flux, and are continually modulated by experiential learning. Traditionally, frequency effects have often been used as diagnostic measures for the existence of stored lexical representations, where surface frequency effects provide evidence for non-compositional, holistic/stored representations of complex words, and stem or suffix (morphemic) frequency effects signify morpheme-specific representations. Under the NDL account, the relatively early surface frequency and stem frequency effects (in eye movements), may be interpreted as a very first stage of the reading process, in which frequency is an approximation of the strength of an orthographic cue for the activation of a meaning in a learning system (for further discussion see Baayen, 2014). Importantly, while vastly different in their implementations and premises, both NDL and multi-route models – unlike the form-then-meaning accounts – allow for even partial orthographic information to activate meanings of complex words and their morphemes, in a parallel rather than sequential way.

While our findings are directly relevant (and problematic) for the behavioral basis of the form-then-meaning account, their relevance for the neurophysiological basis of that account calls for elaboration. As discussed in the Introduction, the causal nature of neural activation for a behavioral effect implies that the earliest detection of the onset of a behavioral effect is an upper temporal bound for the neural instantiation of that effect. Moreover, we reasoned that the behavioral effect cannot exist without its neural instantiation (irrespective of whether that neural instantiation is able to be detected by the particular brain imaging method). It does not, however, require that every effect that is present in the neural record has a behavioral counterpart, nor does it necessitate that a late behavioral effect is also late to emerge in the neural record. This situation is well demonstrated by the effect of word length. It is found to have an early signature in every EEG and MEG study under discussion; yet it has either a relatively late onset in behavioral latencies, or is too weak to produce an appreciable effect (except see Studies 5 and 7 for an early effect of word length in eye movement data). As argued above, on the whole, these findings are not contradictory, nor are they surprising given a body of evidence from the eye tracking literature. While word length is a major determinant of where the eye lands on the word and whether the eye will make more than one fixation on that word, this predictor tends to have little to no effect on the duration of first fixation on the word, i.e. the class of eye movement latency that we analyze in this study. See for instance effect sizes of word length across the eye movement record in Kuperman and Van Dyke (2011b; Figure 6).

Since neural and behavioral effects might be decoupled in their timing (with the latter setting an upper bound for the former), there is a logical possibility that the relative order of processing stages in the brain does in fact follow the sequence suggested by the form-then-meaning account. For this order to reconcile with the order that we observed in our data, the following scenario needs to be in place. First, all effects need to emerge in the brain activity record prior to 140 ms (the point at which earliest onsets are established in the eye movement record). Second, the effects that the form-then-meaning account predicts to emerge in the brain activity last (e.g., surface frequency, semantic transparency or whole-word valence) would need to unfold such that they are the first to affect word recognition behavior, cf. timelines in Figures 2 and 3. Conversely, effects that are expected early in the brain (e.g., morpho-orthographic variables like derivational entropy or lemma transition probability) would also need to elicit a change in behaviour relatively late in a response latency. In this scenario, not only do the derived word and its morphemes need to be fully activated in form and meaning extremely early, but also the predicted order of processing stages would effectively need to be reversed when translating from the brain activation to response latencies. While our present data do not enable us to rule out this possibility, we find that a more parsimonious proposal is to presume a degree of correspondence between the observed relative order of effects in the lexical decision and eye movement records and those in the neurophysiological record, such that earlier/later effects in the former tend to also emerge earlier/later in the latter. In other words, on grounds of parsimony, we presume that the neural record reflects the same form-and-meaning sequence that we find in behavioral data.

Other findings and limitations

An important contribution of this paper is the novel application of the non-parametric distributional survival analysis technique (Reingold & Sheridan, 2014) to the study of cognitive processes involved in complex word recognition. The technique enables researchers to extract temporal information from a distribution of responses, rather than generate a finding based on the central tendency of the final outcome (i.e. the latency of a lexical decision response or eye-fixation). In all of our analyses, all divergence points were detected at timepoints when less than 50% of responses/eye-fixations were terminated, and in a vast majority of cases were shorter than the mean response time in each dataset. In the Methods section, we outlined the complementary nature of survival analysis to methods based on explaining variance (e.g., ANOVA or regression). A contrast between the means of conditions (i.e. a reliable effect in ANOVA or regression) will also show in the divergence between condition-based survival curves around the mean latency. However, the size of the contrast is not inherently related to how early this effect emerges, and so we argued that the estimated divergence points of target effects are not predicted to correlate with the magnitude of those effects. We tested this dissociation by calculating correlations between median divergence points for all predictors that were deemed reliable in the survival analysis (see criteria in the Methods) and the correlations of those predictors with the outcome variable (lexical decision or eye-fixation latency). We found no relationship between effect strength and the ordering of divergence point estimates in any of our studies [Study 1: $\rho = 0.55$; Study 2: $\rho = 0.32$; Study 3: $\rho = 0.07$; Study 4: $\rho = 0.54$; Study 5: $\rho = 0.05$; Study 6: $\rho = 0.51$; Study 7: $\rho = 0.56$; all p-values > 0.05]. This finding suggests that the onsets of effects, as estimated by the survival technique, are not a mere replica of the strength of the variable as a predictor of RTs.

However, there are limitations to the current study. Firstly, the distributional approach we adopt here is currently unable to evaluate the relative importance of a given variable on the survival rate throughout the entire time-course. For example, whereas surface frequency might influence the very earliest responses within a distribution of RTs, we do not know whether surface frequency retains influence for the slowest responses (though see other types of distributional analyses, Balota & Spieler, 1999; Balota & Yap, 2011; Ratcliff, 1979; Staub et al., 2010; Van Zandt, 2002; Vincent, 1912). Secondly, we are also unable to control for idiosyncratic variation that is due to individual differences in participants, or for the items themselves. Secondly, in the current survival implementation, each lexical variable is considered independently of one another. Although we addressed the issue of collinearity (see supplementary materials section S1), there is also a need for future parametric solutions to take into account the combined effect of lexical variables (and also variables pertaining to individual differences in reading skill) on the word recognition time-course. Ultimately, more complex solutions to modelling survival rates of lexical decision responses will be able remedy many of these issues. We also argue that more such distributional analyses may provide a promising tool for comparing behavioural data with the results from data sampled from the continuous electrical or magnetic signal associated with that decision (i.e. MEG and ERP).

Lastly, our use of eye-tracking studies of sentence reading may give rise to criticism. Because words are previewed in the eye's parafovea, some bottom-up information about target derived words may become available to the human processor before the word is fixated and affect the timeline of its processing. There is no consensus at present on whether semantic information can be extracted parafoveally (see review by Schotter, Angele & Rayner, 2012), yet there is robust evidence that orthographic and phonological information can. If so, we would expect formal properties of complex words to get a recognition boost and affect the time-course of complex word processing earlier than they would if not parafoveal preview was available. Largely, this expectation is borne out: effects of formal properties emerge earlier in eye-tracking Studies 5-7 than in lexical decision Studies 1-4. Another expected consequence of word presentation in context is that its semantics would be accessed relatively late, as it does not enjoy the same parafoveal preview advantage as the word's formal features. Since this is contrary to the fact, we suggest that parafoveal preview does not give a spurious boost to the onset of either semantic properties or surface frequency, i.e. the features diagnostic for the form-then-meaning account. Another source of early information about the target word may come from the top-down predictability of the word in its context. All our sentences were constructed to keep the context preceding the word semantically neutral, and Cloze predictability norms (collected in separate experiments) confirmed that the actual target words were essentially unpredictable. We cannot rule out however, that some syntactic information, including expected part of speech, would become available to the reader from the word's context. It is unlikely that such information would point to the target word as a derived word though, and so a decomposition advantage is not expected to arrive from context. Still, we agree that our speculations need to be tested in future eye-tracking studies of isolated word reading. At present, we confine ourselves to reminding the reader that such studies (reviewed in the Introduction) demonstrate incompatibility of the behavioral time window of morphological processing with that falling out of neurophysiological studies.

Finally, an additional contribution of this paper is the addition of valence to the palette of semantic variables with which one can explore the time-course of semantic access during complex word recognition. Our findings corroborate with Kuperman's (2013) finding that the emotional connotation of a compound word and those of its constituents play a role in recognizing the whole word. As well as demonstrating that semantic access to the atomic properties of compound words extends to derived words (see also Feldman, Brown and Pastizzo, 2006, who found effects of the atomic property of stem concreteness in a priming study with morphologically complex primes), these findings also supplement current theoretical treatments of the effects of emotion on word recognition. Irrespective of morphological complexity, a number of experiments have shown that negative words, such as *vomit*, elicit slower responses as compared to neutral stimuli, such as *nunnery* (e.g., Citron, 2012; Wentura, Rothermund, & Bak, 2000; Kuperman, Estes, Brysbaert, & Warriner, 2014). These results are typically attributed to the human organism's *automatic vigilance* to negative stimuli (Erdelyi, 1974). This hypothesis proposes that negative words capture attention for longer and thus generate slower lexical decision responses. The relatively early time at which a divergence between positive and negative stimuli exerts an influence on lexical decision times demonstrates the rapidity of the automatic response to the emotional content of linguistically encoded stimuli. That these effects are early for both simplex and complex words, across Dutch and English, and across two methodologies, demonstrates a reliable and privileged role of valence during word recognition.

Conclusion

Our results demonstrate the utility of advanced distributional analyses for theoretical questions that are critically dependent on a fine-grained determination of the absolute and relative timing of lexical effect onsets. Our findings provide evidence against the predictions of the form-then-meaning account of morphological processing. This evidence is issued against both the relative order of processing stages and the purported absolute timeline of their emergence in the neural and behavioral record. Specifically, the results of survival analyses applied to lexical decision and eye movement data in English and Dutch narrow down the expected timeline of semantic influence on a behavioural response to 140-250 ms, thus setting a temporal upper bound for its respective neural activation to emerge. These results are important as guidelines for an expected timeline of formal and semantic effects across paradigms of morphological processing research. Together with the apparent relative ordering of lexical effects, the observed reductions to the absolute timeline of effect onsets render our results convergent with the premises of form-and-meaning models of complex word recognition.

Appendix A: Descriptive statistics for lexical variables

Table A1: Descriptive statistics for lexical variables used in studies 1-4.

Study		RT	Surface freq.	Stem freq.	Valence of word	Valence of stem	LSA distance	Derivational Entropy	Suffix prod. (SP)	Lemma transition probability (TPL)	Word length	Orthographic neighbourhood density (OLD20)	Bigram transition probability (TPB)	Form suffix prod. FSP)
	Min	236	5	31	1.85	1.53	0.296	0	5	0.117	4	1.000	0.0000	88
Study 1	1st Qu.	488	160	1878	3.90	4.50	0.572	0.8324	1214	0.677	6	1.400	0.0013	2056
Derived	Median	558	352	9019	5.00	5.59	0.666	1.309	2276	1.000	6	1.650	0.0026	6646
	Mean	606	1362	32630	4.95	5.37	0.663	1.377	2276	0.834	6.59	1.732	0.0045	6448
	3rd Qu.	666	956	24990	5.86	6.47	0.757	1.93	3499	1.000	7	1.925	0.0062	13050
	Max	2213	89540	681500	8.21	8.21	0.995	3.857	3499	1.000	11	3.700	0.0495	13050
	Min	347	6	8	1.67	1.38	0.368	0	56	0.400	5	1.100	0.0000	3
Study 2	1st Qu.	490	76	420	2.67	2.88	0.512	1.059	501.8	0.685	6	1.850	0.0016	41
Derived (Dutch)	Median	548	291	1918	4.92	4.80	0.623	1.713	2991	0.800	7	2.100	0.0025	314
	Mean	578	983	6102	4.34	4.30	0.642	1.985	3187	0.793	7.39	2.195	0.0056	575
	3rd Qu.	632	882	4772	5.75	5.50	0.756	2.919	3097	0.952	8	2.600	0.0049	335
	Max	1315	15500	139900	6.42	6.42	1.022	5.289	10330	1.000	11	3.700	0.0683	5508
	Min	281	26	116	1.90	2.65	0.549	0	24	0.147	4	1.000	0.0002	225
Study 3	1st Qu.	477	336	606	4.49	4.61	0.809	0	24	0.691	5	1.112	0.0023	833
Pseudo-derived	Median	547	1291	2246	5.38	5.36	0.863	0.2476	2276	0.863	6	1.475	0.0046	6646
	Mean	591	6460	13900	5.22	5.19	0.856	0.6655	1829	0.805	5.82	1.452	0.0074	5737
	3rd Qu.	648	5585	10020	5.99	5.82	0.915	1.248	3499	0.997	6	1.700	0.0094	6646
	Max	1877	73400	362900	7.71	7.60	1.008	3.658	6254	1.000	9	2.600	0.0495	13050
	Min	272	9	25	1.95	2.12	0.361			0.185	5	1.000	0.0001	7
Study 4	1st Qu.	484	189	532	4.25	4.25	0.807			0.617	6	1.500	0.0032	416
Form controls	Median	552	434	1851	5.30	5.21	0.873			0.841	6	1.725	0.0054	983
	Mean	600	2599	18220	5.22	5.15	0.858			0.775	6.19	1.687	0.0112	2054
	3rd Qu.	660	2154	13210	6.12	5.99	0.932			0.989	6	1.850	0.0097	3098
	Max	2328	28430	717100	8.05	8.34	1.079			1.000	9	3.050	0.2975	18170

Table A2: Descriptive statistics for lexical variables used in studies 5-7.

Study		First fixation duration	Surface freq.	Stem freq.	Valence of word	Valence of stem	LSA distance	Derivational Entropy	Suffix prod. (SP)	Lemma transition probability (TPL)	Word length	Orthographic neighbourhood density (OLD20)	Bigram transition probability (TPB)	Form suffix prod. FSP)
Study 5	Min	82	7	34	1.85	1.53	0.000	-0.000	436	0.200	5	1.000	0.0001	561
	1st Qu.	181	64	188	4.85	5.05	0.493	0.849	653	0.742	7	1.800	0.0010	1148
	Median	220	142	815	5.42	5.80	0.614	1.265	653	0.891	9	2.550	0.0026	1148
	Mean	231	306	5803	5.37	5.60	0.600	1.330	1669	0.831	9	2.413	0.0049	2911
	3rd Qu.	266	322	4595	6.30	6.65	0.710	1.850	3499	0.981	11	2.850	0.0076	6646
	Max	946	3206	203900	8.60	8.78	0.985	3.244	3499	1.000	13	4.100	0.0220	6646
Study 6	Min	82	4	20	1.90	1.55	-0.000	-0.000	370	0.149	6	1.700	0.0003	225
	1st Qu.	177	40	290	4.29	4.42	0.522	0.847	370	1.000	8	2.450	0.0010	225
	Median	215	97	813	5.90	5.75	0.618	1.346	493	1.000	9	2.700	0.0030	538
	Mean	225	412	4864	5.61	5.49	0.626	1.400	517	0.960	9	2.711	0.0050	591
	3rd Qu.	259	216	2477	7.00	6.70	0.709	1.930	607	1.000	10	2.950	0.0093	800
	Max	881	14270	203900	8.16	8.37	0.943	3.315	633	1.000	12	4.250	0.0203	886
Study 7	Min	81	4	20	1.90	1.55	-0.000	-0.000	370	0.149	6	1.700	0.0003	225
	1st Qu.	177	40	290	4.29	4.42	0.522	0.847	370	1.000	8	2.450	0.0010	225
	Median	215	97	813	5.90	5.75	0.618	1.346	493	1.000	9	2.700	0.0030	538
	Mean	233	412	4864	5.61	5.49	0.626	1.400	517	0.960	9	2.711	0.0050	591
	3rd Qu.	268	216	2477	7.00	6.70	0.709	1.930	607	1.000	10	2.950	0.0093	800
	Max	816	14270	203900	8.16	8.37	0.943	3.315	633	1.000	12	4.250	0.0203	886

Appendix B: Correlation matrices for lexical variables

Table B1: Correlation matrix of lexical variables and lexical decision reaction times in the English derived words condition (Study 1). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Surface freq.												
2. Stem freq.	0.28***											
3. Valence word	0.20***	0.21***										
4. Valence stem	0.15***	0.41***	0.57***									
5. LSA distance	-0.25***	0.08***	-0.08***	-0.03***								
6. Deriv. entropy	0.05***	0.47***	0.14***	0.21***	0.07***							
7. SP	0.01	-0.16***	0.07***	-0.05***	-0.02**	-0.18***						
8. TPL	0.07***	0.01	-0.11***	0.02*	0.07***	-0.11***	-0.55***					
9. Word length	-0.14***	0.13***	-0.04***	0.07***	0.04***	0.20***	-0.33***	0.03***				
10. OLD20	-0.09***	0.07***	-0.05***	0.09***	-0.02*	0.14***	-0.58***	0.34***	0.77***			
11. TPB	0.02**	-0.01	-0.09***	-0.05***	0.02*	-0.04***	0.16***	-0.06***	-0.03***	-0.19***		
12. FSP	0.08***	-0.32***	-0.02**	-0.10***	-0.02***	-0.33***	0.46***	0.00	-0.68***	-0.61***	0.20***	
13. Reaction time	-0.21***	-0.09***	-0.05***	-0.04***	0.08***	-0.04***	0.03***	-0.02**	0.01	-0.01	0.01	0.02*

Table B2: Correlation matrix of lexical variables and reaction times in the Dutch derived words condition (Study 2). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Surface freq.												
2. Stem freq.	0.21***											
3. Valence word	0.22***	0.22***										
4. Valence stem	0.24***	0.15***	0.82***									
5. LSA distance	-0.38***	-0.07***	0.10***	-0.06**								
6. Deriv. entropy	-0.09***	0.32***	-0.07***	0.00	0.14***							
7. SP	0.25***	-0.10***	-0.09***	0.00	-0.21***	0.07***						
8. TPL	0.21***	0.29***	0.06**	0.09***	-0.27***	0.05**	0.15***					
9. Word length	-0.16***	0.08***	-0.05*	-0.07***	-0.07***	-0.11***	-0.28***	0.17***				
10. OLD20	-0.23***	0.02	0.05**	-0.02	0.03	-0.12***	-0.61***	0.05*	0.69***			
11. TPB	0.03	-0.12***	0.01	0.05**	0.14***	-0.02	0.12***	-0.16***	-0.11***	-0.26***		
12. FSP	0.18***	-0.23***	-0.15***	-0.06***	-0.14***	-0.09***	0.79***	-0.05**	-0.53***	-0.77***	0.25***	
13. Reaction time	-0.15***	-0.10***	-0.07***	-0.06***	0.04*	-0.07***	-0.01	-0.04*	0.05**	0.08***	-0.02	-0.02

Table B3: Correlation matrix of lexical variables and reaction times in the English pseudo-derived words condition (Study 3). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Surface freq.												
2. Stem freq.	0.01											
3. Valence word	0.43***	0.10***										
4. Valence stem	-0.01	0.37***	0.20***									
5. LSA distance	-0.09***	-0.43***	0.00	-0.07***								
6. Deriv. entropy	0.01	0.37***	0.05**	0.08***	-0.18***							
7. SP	0.17***	-0.08***	0.20***	0.18***	0.09***	-0.06***						
8. TPL	0.16***	-0.03*	0.05***	-0.12***	-0.07***	0.04*	-0.12***					
9. Word length	-0.01	-0.03	0.20***	-0.04**	-0.13***	0.15***	-0.01	-0.05**				
10. OLD20	-0.03*	0.09***	0.11***	-0.04**	-0.20***	0.12***	-0.26***	-0.07***	0.69***			
11. TPB	-0.05**	0.14***	0.05***	0.10***	0.04**	-0.12***	0.04**	-0.05***	0.20***	-0.20***		
12. FSP	0.17***	0.01	0.03	0.06***	-0.02	-0.07***	0.63***	0.14***	-0.43***	-0.43***	-0.02	
13. Reaction time	-0.26***	-0.03	-0.15***	0.04**	0.09***	0.00	-0.01	-0.01	-0.02	-0.04**	0.05**	-0.02

Table B4: Correlation matrix of lexical variables and reaction times in the English form control condition (Study 4). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
11.	12.									
1. Surface freq.										
2. Stem freq.	0.02									
3. Valence word	0.49***	0.02								
4. Valence stem	0.05***	0.47***	0.08***							
5. LSA distance	-0.09***	-0.21***	-0.18***	-0.10***						
6. TPL	0.20***	-0.08***	0.12***	-0.01	0.11***					
7. Word length	-0.11***	0.01	-0.16***	-0.03*	0.05***	0.06***				
8. OLD20	-0.11***	0.02	-0.11***	0.12***	0.18***	0.11***	0.52***			
9. TPB	0.04**	-0.02	0.00	0.03*	0.22***	0.07***	-0.04**	-0.11***		
10. FSP	0.04**	-0.12***	0.07***	-0.18***	-0.15***	-0.20***	-0.22***	-0.43***	-0.24***	
11. Reaction time	-0.28***	-0.01	-0.16***	-0.02	0.07***	0.02	0.05***	0.07***	0.01	-0.05***

Table B5: Correlation matrix of lexical variables and first fixation durations in Study 5 (Experimental list 1). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Surface freq.												
2. Stem freq.	0.22***											
3. Valence word	-0.02	0.08***										
4. Valence stem	0.01	0.30***	0.73***									
5. LSA distance	-0.19***	0.15***	0.06**	0.16***								
6. Deriv. entropy	0.03	0.24***	0.06**	0.07***	0.28***							
7. SP	-0.04	0.31***	-0.07***	0.00	-0.06**	-0.05*						
8. TPL	-0.03	0.09***	0.08***	0.01	-0.22***	0.05*	-0.11***					
9. Word length	-0.14***	-0.48***	0.02	-0.10***	0.01	0.07***	-0.65***	0.17***				
10. OLD20	-0.12***	-0.47***	0.08***	-0.01	0.07***	0.04*	-0.75***	0.14***	0.89***			
11. TPB	0.05*	0.05**	-0.03	0.00	-0.14***	-0.16***	0.34***	-0.08***	-0.21***	-0.23***		
12. FSP	-0.04	0.31***	-0.07***	0.00	-0.06**	-0.05*	1.00***	-0.11***	-0.65***	-0.75***	0.34***	
13. Fixation duration	-0.07***	-0.08***	0.02	-0.02	0.01	0.03	-0.01	0.03	0.10***	0.06**	0.05*	-0.01

Table B6: Correlation matrix of lexical variables and first fixation durations in Study 6 (Experimental list 2). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Surface freq.												
2. Stem freq.	0.17***											
3. Valence word	0.26***	0.23***										
4. Valence stem	0.25***	0.33***	0.86***									
5. LSA distance	-0.19***	0.05**	-0.11***	-0.13***								
6. Deriv. entropy	0.09***	0.11***	0.14***	0.18***	0.17***							
7. SP	-0.08***	-0.08***	0.09***	0.13***	0.14***	0.16***						
8. TPL	-0.10***	0.08***	0.12***	0.06***	-0.05**	-0.15***	-0.22***					
9. Word length	-0.04**	-0.15***	0.13***	0.08***	0.12***	0.03	0.29***	0.10***				
10. OLD20	-0.21***	0.01	0.11***	0.05**	-0.08***	-0.14***	-0.02	0.21***	0.61***			
11. TPB	-0.02	0.05**	-0.12***	-0.08***	0.06***	0.01	0.01	0.09***	0.11***	-0.02		
12. FSP	-0.08***	-0.08***	0.09***	0.13***	0.14***	0.16***	1.00***	-0.22***	0.29***	-0.02	0.01	
13. Fixation duration	-0.07***	-0.06***	-0.06***	-0.05***	0.02	-0.01	0.02	-0.01	0.00	-0.01	0.04*	0.02

Table B7: Correlation matrix of lexical variables and first fixation durations in Study 7 (Experimental list 2). The lower triangle provides Spearman correlation coefficients. ***Correlation is significant at the .001 level. **Correlation is significant at the .01 level. *Correlation is significant at the .05 level.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Surface freq.												
2. Stem freq.	0.18***											
3. Valence word	0.27***	0.22***										
4. Valence stem	0.25***	0.33***	0.86***									
5. LSA distance	-0.20***	0.04*	-0.13***	-0.14***								
6. Deriv. entropy	0.07***	0.12***	0.15***	0.18***	0.17***							
7. SP	-0.07***	-0.09***	0.08***	0.10***	0.15***	0.16***						
8. TPL	-0.09***	0.05***	0.12***	0.05***	-0.04*	-0.15***	-0.23***					
9. Word length	-0.01	-0.17***	0.10***	0.05***	0.13***	0.02	0.29***	0.10***				
10. OLD20	-0.18***	0.00	0.09***	0.04*	-0.08***	-0.14***	-0.05**	0.21***	0.62***			
11. TPB	0.01	0.05**	-0.14***	-0.10***	0.05***	0.02	0.03	0.10***	0.11***	-0.02		
12. FSP	-0.07***	-0.09***	0.08***	0.10***	0.15***	0.16***	1.00***	-0.23***	0.29***	-0.05**	0.03	
13. Fixation duration	-0.10***	-0.04**	-0.05**	-0.04**	0.03	0.00	0.02	-0.02	0.02	0.00	0.01	0.02