



# A tipping point in word recognition? Investigating the relationship between root and form frequency across visual and auditory modalities

Hanno Müller<sup>1,2</sup> · Louis ten Bosch<sup>1</sup> · Mirjam Ernestus<sup>1</sup>

Received: 7 April 2024 / Accepted: 16 September 2025  
© The Author(s) 2025

## Abstract

For various theories of human word recognition, the question of how the recognition of suffixed words is influenced by the morphological root or the surface form of the word is of considerable relevance. According to many theories (e.g., Baayen et al., 1997b), the morphological root predominantly guides the recognition process unless the word is of a (relatively) high frequency of occurrence. We tested this ‘tipping point’ hypothesis by comparing a statistical model based on this hypothesis with two alternative statistical models: one assuming that word recognition is always root-driven and another assuming it is always form-driven. To this end, we modeled response time distributions from two large-scale lexical decision experiments in Dutch – one visual and one auditory – focusing on three suffixes: the plural suffix *-en* for nouns, the derivational suffix *-heid* for nominalisations, and *-t* as the second/third person singular present tense suffix for verbs. Our results indicate that words with the suffixes *-t* and *-heid* are retrieved as whole forms in both visual and auditory word recognition. In contrast, words with the suffix *-en* are best accounted for by both the root-driven and the form-driven models in auditory word recognition, while in visual word recognition, they support the tipping point hypothesis. Taken together, our findings suggest that both root-driven and form-driven principles are relevant for word recognition, while the assumption of a categorical tipping point is less tenable. This study contributes to our understanding of word recognition mechanisms in both localist and distributional-connectionist theoretical frameworks.

**Keywords** Morphological processing · Auditory word recognition · Visual word recognition · Computational modelling · Lexical decision · Tipping point

---

✉ H. Müller  
[hanno.mueller@ru.nl](mailto:hanno.mueller@ru.nl)

<sup>1</sup> Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

<sup>2</sup> English Language and Linguistics, Heinrich-Heine-University, Düsseldorf, Germany

## 1 Introduction

In normal and healthy conditions, we effortlessly process spoken words and, once we have learned to read, written words as well. Yet the mechanisms underlying word recognition are far from trivial and have been the subject of numerous studies for several decades (e.g., Morton, 1969; Grainger & Jacobs, 1996; Chuang & Baayen, 2021). A particular focus of research on word recognition has been on how we recognise morphologically complex words (for overviews see Amenta & Crepaldi, 2012; Milin et al., 2018; Stevens & Plaut, 2022), that is, words that can be divided into smaller meaningful units (i.e., morphemes). For example, *walked* is complex because it can be broken up into the root *walk* and the suffix *-ed*. The present study contributes to research on how humans recognise morphologically complex words by investigating the roles of the frequencies of both the complex words (e.g., *walked*) and their roots (e.g., *walk*).

There are strong indications that the recognition of a morphologically complex word (henceforth: complex word) is affected by the word's morphological structure and morphologically related words. For instance, words (e.g., *walk*) are recognised faster when more morphologically related words (e.g., *catwalk*, *walked*, *walkie-talkie*, *walkman*) exist in the language, an effect known as the morphological family size effect (e.g., Schreuder & Baayen, 1997; De Jong, 2002; Perdijsk et al., 2012). To give another example, the short presentation of the root of a morphologically complex word before the presentation of the complex word (e.g., *walk* – *walked*) leads to a faster recognition of the complex word than when a morphologically unrelated word is presented beforehand (e.g., *talk* – *walked*; Geary & Ussishkin, 2018; De Grauwe et al., 2019). Vice versa, the short presentation of a complex word before the presentation of this word's root (e.g., *walked* – *walk*) leads to a faster recognition of this root compared to when the root is preceded by a morphologically unrelated word (e.g., *talked* – *walk*; Beyersmann et al., 2016; Wilder et al., 2019). Importantly, such priming effects are significantly stronger for primes that are morphologically related to the target (e.g., *walked* – *walk*) compared to primes that are solely orthographically related to the target (e.g., *department* – *depart*), which indicates that the existence of a morphological relationship increases priming beyond what is explained by a mere orthographic relationship (e.g., Rastle et al., 2000; Cho et al., 2024).

This study focusses on the cumulative root frequency (henceforth root frequency), another measure based on morphological structure. The root frequency is the total frequency of all words that contain the presented word's root. Words are recognised quicker, the higher their root frequencies (henceforth *root frequency effect*, e.g., Taft, 1979; Meunier & Segui, 1999; Solomyak & Marantz, 2010; Sánchez-Gutiérrez et al., 2018). Traditionally, it has been assumed that the root frequency effect occurs independently of the often-replicated effect of form frequency (e.g., Taft, 1979; Carrazza et al., 1988; Colé et al., 1989; Alegre & Gordon, 1999), which is based on the number of occurrences of the word's surface form (e.g., Brysbaert et al., 2018). The assumption of independent root and form frequency effects is challenged, though, by studies suggesting an interaction between the two effects (e.g., Baayen et al., 2007; Luke & Christianson, 2011). For instance, Baayen et al. (2007) reported that, for English written words, higher root frequencies result in slower recognition times in high form frequency words and in quicker recognition times in low form frequency words.

## 1.1 Theoretical explanations for the root frequency effect (in interaction with form frequency)

Theories of how humans recognise morphologically complex words can be broadly categorised based on their view of the mental lexicon: either a localist view or a distributional-connectionist view. The localist view (e.g., Morton 1969, 1970; Jackendoff & Audring, 2020; Taft, 2023) assumes that different linguistic units (e.g., morphemes, lemmas, words) have their own representations in the mental lexicon. The root frequency effect is the result of the involvement of the word's root in the recognition of the word, either because regular morphologically complex words are not stored themselves in the mental lexicon (e.g., Taft & Forster, 1975), or because morphologically complex words can be recognised both on the basis of their root and their own lexical representations (e.g., Baayen et al., 1997b; Jackendoff & Audring, 2020; Taft, 2023), or because the lexical representations of complex words refer to the words' root (e.g., Meunier & Segui, 1999). The interaction of root frequency and form frequency can well be accounted for by theories that assume that both a word's own lexical representation and that of its root can be involved in word recognition. A word's root may play a more important role when the representation of the word itself is difficult to access, for instance because of its low frequency of occurrence (Baayen et al., 2007). In contrast, when the representation of the word is easy to access, the representation of the root may hinder word recognition because it activates morphologically related words as competitors for the word to be recognised.

According to the distributional-connectionist view (e.g., Seidenberg & McClelland, 1989; Gonnerman et al., 2007; Blevins, 2016), the recognition of words does not involve subunit representations, and some distributional-connectionist models do not even assume lexical representations for words (Chuang & Baayen, 2021). Instead, word recognition is conceptualised as the mapping of an incoming signal (e.g., a letter sequence or an audio signal) onto a meaning representation. This mapping is often implemented in artificial neural networks that are trained with pairs of word and corresponding meaning representations. During training, given an input word representation, the network predicts a meaning representation, compares this prediction with the actual meaning representation, and then uses the difference between predicted and observed meaning to adjust itself so that it yields more accurate predictions when it encounters the same input the next time. In such a network, the root frequency effect emerges because root morphemes tend to have consistent form-meaning-relationships. For instance, the words *catwalk*, *walked*, and *walkman* all contain the sequence *walk* and a meaning that is related to 'moving around'. In distributional-connectionist models, it is assumed that the most influential weights for the recognition of high frequency words are fine-tuned to these words themselves, whereas the most influential weights for the recognition of low frequency words are fine-tuned to these words' roots (i.e., more often encountered items are learned better; Seidenberg & McClelland, 1989). In order to be able to further develop both the localist and the distributional-connectionist theories, more insight is necessary in how root frequency affects word recognition, as a function of form frequency. Differences may be expected between visual and auditory word recognition since they differ from each other in crucial aspects.

## 1.2 Visual versus auditory word recognition

As mentioned above, morphological structure affects word recognition both in the written and the auditory modality. Nevertheless, we may expect differences between the two modalities. The fact that most written words can be perceived in their entire forms in one glance, while spoken words unfold over time, has two implications.

First, a written word's root can often be perceived immediately upon stimulus presentation, whereas a spoken word's root is always perceived after its prefix and before its suffix. As a consequence, a word's root may play a larger role in the recognition of suffixed words than of prefixed words in the auditory modality, which is supported by the study by Müller et al. (2024) in the context of family size effects.

Second, whereas in the visual modality, the final segment of the word can be immediately perceived, in the auditory modality, listeners have to hear the complete word before they can hear the final segment. This affects how participants react in the many lexical decision experiments that have been conducted to investigate the role of a word's morphological structure, and in which the last segment of a stimulus can turn the stimulus from a real word into a pseudoword. Listeners tend to wait until word offset to make their lexicality decision (e.g., Ernestus & Cutler, 2015) and word duration is the strongest predictor of auditory lexical decision reaction times (RTs). In contrast, the most important predictor for visual lexical decision RTs is not word length in number of characters – the visual counterpart to word duration – but form frequency (e.g., Ferrand et al., 2018).

Another difference between written and spoken words is that, in many languages, the spelling of a root (e.g., *walk*) is typically independent of whether it is embedded in a morphologically complex word (e.g., *walked*). Nevertheless, the pronunciation of a root may depend on whether it is embedded in a complex word and in which word. To mention just a small possible difference, in spoken words of languages like English and Dutch, roots that are realised in isolation are longer than roots realised in complex words (e.g., Baayen et al., 2003). A more pronounced difference can be found in the vowel (of the root) in the word pair *breath* /brɛθ/ versus *breathing* /bri:ðɪŋ/. These differences may make it more difficult to recognise a given root in different words in auditory word recognition, which may decrease the role of the frequency of the root in auditory word recognition.

How much the pronunciation of a root in a word differs from its most common pronunciation may vary with the affixes in the word. For instance, in Dutch, root-final obstruents may differ in their voicing depending on whether they are in syllable final position (together with a consonant initial affix) or in syllable-initial position, as a consequence of resyllabification with a following vowel initial affix.

## 2 Goal of the present study

The present study investigates the role of root frequency in the recognition of morphologically complex words, in both the written and the auditory modality, as a function of the suffix. Specifically, we compare the role of root frequency with the role of word form frequency. We do so in two ways. On the one hand, we compare a model

of word recognition that only takes root frequency into account with a model that only takes word form frequency into account. On the other hand, we test the meta model (see Sect. 3), which assigns an important role to the relation between the two frequencies, assuming that root frequency is the only important predictor for words with relatively high root frequencies, and that word form frequency is the only important predictor for words with relatively high word form frequencies. We will refer below to this type of model as a model with a tipping point.

We analyse RTs from existing lexical decision data in Dutch. We chose Dutch because the meta model was already applied to Dutch (Baayen et al., 1997b) and because of the availability of large datasets of both visual and auditory lexical decision data for this language. Similarly large datasets are also available for English (Keuleers et al., 2012; Tucker et al., 2019) and French (Ferrand et al., 2010, 2018), but we preferred Dutch, because it has a richer morphological system than English, and because it has a more transparent orthography-pronunciation relationship than French and English, which makes word recognition in the visual modality more similar to word recognition in the auditory modality.

We focus on words with three different suffixes (plural *-en*, verbal *-t*, and *-heid*), which differ in characteristics that may affect the hypothesised tipping point. This will shed light on how the characteristics of the affixes affect the word recognition process.

### 3 The tipping point model tested in the present study

We will test the role of the tipping point of root and form frequency on the basis of a computational model that was especially popular at the end of the last century, the meta model (Schreuder & Baayen, 1995; Baayen et al., 1997b). Despite its publication date of about 30 years ago, throughout the years, various authors have drawn on the meta model to explain their experimental findings on, among others, the processing of nouns in Italian (Baayen et al., 1997a), Hebrew (Vaknin-Nusbaum & Shimron, 2011; Vaknin-Nusbaum, 2025), Russian (Savinova & Malyutina, 2021), Dutch and German (Reifegerste et al., 2017), compounds in Finnish (Pollatsek & Hyönä, 2005) and Chinese (Zou et al., 2023), and English affixed free and bound roots (Solomyak & Marantz, 2009) as well as English suffixed words (Dawson et al., 2018).

The meta model was developed to predict RTs to simplex (i.e., root words) and complex words such as plurals in Dutch and English. It assumes that recognition times consist of multiple time components. The first, constant, components are an *initial mapping time* for the mapping of the signal onto lexical representations (for simplex and complex words) and an *execution time* required to carry out a response. The third component is an *activation time*, the time it takes until a lexical representation reaches threshold activation level. For simplex words, the recognition of which is assumed to consistently be root-driven, the activation time is inversely proportional to the word's cumulative stem frequency. Baayen et al. (1997b) define cumulative stem frequency as the “summed frequencies of a stem and all words in which that stem occurs”. A complex word's activation time would be inversely proportional to the word's form frequency in case of form-driven word recognition and to the word's

cumulative stem frequency in case of root-driven word recognition. Above that, root-driven recognition would require additional processing time for the activation of the complex word's representation based on the activations of the root and morphologically related words. This time is called the *parsing penalty*. Note that such a parsing penalty would be equivalent to the process of deriving a complex word's meaning from similar word forms by means of analogy in distributional-connectionist models (e.g., Blevins, 2016).

The meta model assumes that root- and form-driven word recognition compete, and that the ratio of the two corresponding frequencies determines how complex words are recognised. When cumulative stem frequency is substantially higher than form frequency, the cumulative stem frequency effect is more likely to outweigh the parsing penalty and shift the recognition process toward root-driven processing. The meta model in its original form (Baayen et al., 1997b) does not predict differences among words that only differ in their affixes. As mentioned above, however, root recognition may be more difficult before certain suffixes than before others, which motivates our investigation of words with different suffixes.

#### 4 Algebraic formulation of the tested models

Our statistical models – one only based on root-driven processing, one only based on form-driven processing, and one with a tipping point – follow Formulae (1), (2) and (3). They are grounded in the principles of mixed-effects modelling (e.g., Baayen et al., 2008) and predict RTs for morphologically simple and complex words using different principles. All formulae contain a by-participant random intercept ( $u_j$ ) and make the RTs depend on characteristics of the word or the experiment that are known to affect word recognition times (e.g., word duration;  $\beta_{1:n}x_{1:n}$ , see below for which characteristics we implemented).

The predictors in addition to root and form frequency were not incorporated in the original 1997 formula of the meta model, and our formulation of the meta model thus deviates from the original formulation. We nevertheless incorporated them because they substantially improve the fit with the experimental data. Baayen et al. (1997b) meticulously controlled their stimuli for properties such as length in letters. We aimed to investigate the processing of a larger variety of stimuli, which better approximates processing of everyday language. We analysed RTs to stimuli that vary in length/duration, for instance.

Formula (1) shows the general specification of the root-driven model. It predicts each RT with an intercept ( $\beta_0$ ) and with the root frequency ( $\beta_{root}F_{root}$ ). In addition, the model assumes a parsing penalty ( $\Delta p$ ), which covers the delay induced by recognising a complex word via its root (and morphologically related words). Formula (2) represents the general specification of the form-driven model. It differs from the root-driven model in two respects. First, it is the form frequency ( $\beta_{form}F_{form}$ ), rather than the root frequency, that contributes to the prediction of RTs, and second, it does not have a parsing penalty. Formula (3) provides the general specification for the tipping point model. It combines (1) and (2) and states that the faster – root- or form-driven

– recognition determines the predicted RTs. See Appendix A for more information. We based our frequencies on the CELEX lexical database (Baayen et al., 1996).

$$RT = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + u_j + \beta_{root} F_{root} + \Delta p, \quad (1)$$

$$RT = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + u_j + \beta_{form} F_{form}, \quad (2)$$

$$RT = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + u_j + \min(\beta_{form} F_{form}, \beta_{root} F_{root} + \Delta p). \quad (3)$$

Unlike what may be suggested by Baayen et al. (1997b), the term *stem* is ambiguous in *cumulative stem frequency* because it is used in the literature to refer to inflectional bases or bound roots, among others (cf. Bauer et al., 2013). In this study, we substitute cumulative stem frequency with root frequency. For computing root frequency, we sum the frequencies of the root and of all words with that root and maximally one other morpheme. For illustration, the root frequency of *spoons* is the sum of the frequencies of *spoon*, *spooned*, *spoonerism*, *spoonful*, *spoons*, and *tablespoon*, but not of the frequencies of *tablespoonful*, *cooking spoons*, or *spoon-feeding*, which consist of more than two morphemes. We did not include words with more than two morphemes in the root frequency computation because many words consisting of more than two morphemes have a frequency of zero or one in 42 million word tokens (Baayen et al., 1996), and therefore are likely not stored in every individual listener's mental lexicon. Because of the low frequencies of words consisting of more than two morphemes, for our six datasets (see Sect. 5) the correlations between the root frequency based on words consisting of at most two morphemes and the root frequency based on all words ranges between 0.94 and 0.99.

In the modelling, we include the two most important characteristics of the words/experiment that predict RTs for lexical decision data. Our first control variable is the word's *Duration* in ms for spoken words (Ernestus & Cutler, 2015) and the word's *Length* in number of letters for written words (Keuleers et al., 2010), because it takes longer to respond to longer than to shorter words. Our second control variable is the *Trial Number* in each session (e.g., Ernestus & Cutler, 2015). This captures participants' gradual adaptation to the task throughout the experiment.

Appendix B shows that the three models as implemented with the formula (1), (2) or (3), respectively, can be correctly distinguished on the basis of a dataset of lexical decision times. It thus forms a proof a concept.

## 5 Outline of six studies

We present six studies, each focusing on a single suffix in either the visual or the auditory modality. Rather than analysing all data with a single statistical model with categorical predictors for suffix (*-en*, *-t*, *-heid*) and modality (*visual*, *auditory*), we deliberately analysed each dataset separately with independent statistical models. This decision was motivated by the fact that different theoretical processing models require different statistical formulations (see formulae (1)–(3)). A unified model with categorical predictors would only identify the best-fitting model across all suffixes and modalities, masking variation in model performance across conditions. In contrast, separately analysing each dataset allows for testing, for example, whether the

tipping point model best predicts RTs for one suffix in one modality, while a different model better predicts RTs for another suffix or modality. This approach enables us to test whether the presence of a tipping point depends on the suffix and/or the modality.

In Study 1, we studied written plural nouns that form their plural with the suffix *-en*. This suffix, which was also investigated by Baayen et al. (1997b), is very productive but can be easily confused with the infinitive suffix *-en*, the plural present tense verb suffix *-en*, and the inflectional suffix *-e* for adjectives, because these affixes are often homonymous. The suffix *-en* forms a syllable with the preceding consonants, which implies a difference in syllable structure between the root presented in isolation and the root in the inflected form. We expect that the tipping point model best predicts the RTs of the *-en* data, because this would be in line with Baayen, Dijkstra, and Schreuder's (1997b) findings. In Study 2, we studied the same type of words in the auditory modality. We expect that the tipping point model best predicts the RTs again, because this would be in line with Baayen et al.'s (2003) findings.

In Studies 3 and 4, we studied second/third person singular present tense verb forms with the suffix *-t*, in the written and auditory modality, respectively. The suffix *-t* differs from the noun plural suffix *-en* (Studies 1 and 2) in several ways. First, while *-en* has multiple functions, *-t* predominantly marks the second/third person singular present tense. Second, while, phonologically, *-en* alters the syllable structure, *-t* does not. Finally, acoustically, *-t* may be more salient than *-en*, but both can be reduced.

Taken together, factors that may favour root-driven processing are present for both *-t* (no frequent homonyms with other morphological functions and, possibly, acoustic saliency) and *-en* (phonological saliency due to changes in the syllable structure). Because of this, differences between *-t* and *-en* are expected, but it is difficult to predict which of the two suffixes is more likely to induce root-driven recognition and thus less likely to provide evidence of a tipping point in the role of form frequency.

Finally, in Studies 5 and 6, we studied nouns that are formed with the suffix *-heid*, again in the visual (Study 5) and the auditory (Study 6) modality. The suffix *-heid* is the longest of all suffixes (both in terms of number of letters and phonemes); it does not have homonyms; it always forms a whole syllable on its own; it is always stressed; and it is the only derivational suffix among the suffixes under scrutiny and therefore it is the only suffix that adds an independent meaning to its base. Because of these differences, we expect words that are formed with *-heid* to have a greater likelihood to undergo root-driven processing (see also Baayen & Neijt, 1997) than words formed with *-t* or *-en* and therefore to be less likely to show evidence for a tipping point.

We based the studies on the visual modality on data from the Dutch Lexicon Project (DLP; Keuleers et al., 2010). DLP contains the responses and RTs from 39 native speakers of Dutch to 14,089 written Dutch content words and 14,089 written pseudowords. The words differ in word class and position of stress, among other characteristics. The pseudowords conform to the phonotactic rules of Dutch and are similar to the words in terms of word length and morphological structure.

We based the studies on the auditory modality on data from the Biggest Auditory Lexical Decision Experiment Yet (BALDEY; Ernestus & Cutler, 2015), representing an extensive auditory lexical decision experiment that was conducted in Dutch. BALDEY contains the responses and RTs from 20 native speakers of Dutch to 2,780

spoken Dutch content words and 2,761 pseudowords. The words systematically differ in word class, position of stress, number of syllables, and morphological structure. Each pseudoword was created by substituting one or two segments of a real word to make sure that the morphological and phonological structure was balanced across the word and pseudoword sets.

## 6 Methods

The data, the scripts used in this paper, and each model's implementation can be found at <https://doi.org/10.34973/jm3a-vj10>.

### 6.1 Data

We separately analysed RTs of subsets of DLP (Keuleers et al., 2010) and BALDEY (Ernestus & Cutler, 2015). These subsets were not controlled for any characteristics of the stimuli such as number of letters or duration in ms, or lemma, bigram, root, or form frequency.

To ensure that we restricted our studies to trials in which participants had processed the stimuli correctly, we only analysed RTs of correct responses that were not given earlier than 100 ms after stimulus onset in DLP and after stimulus offset in BALDEY. We chose these thresholds because the fastest human reaction time is assumed to be around 100 ms (Miller, 1968; Pain & Hibbs, 2007). In DLP, stimuli are presented at once and can thus be recognised as words on stimulus presentation (and the fastest possible RT is thus 100 ms after stimulus onset). In BALDEY, stimuli are incrementally presented due to their auditory nature, and the last speech sound segment could turn a word into a pseudoword. Thus, in BALDEY, stimuli can be recognised as words at its earliest at word offset (and the fastest possible RT is thus 100 ms after stimulus offset).

#### 6.1.1 Plural nouns ending in *-en*

In Studies 1 and 2, in order to better estimate the effects of the control variables in the statistical model, we studied responses to both plural and singular nouns, although for investigating the tipping point hypothesis, studying plural nouns would be sufficient. We selected both the singular nouns in the visual and auditory datasets that form their plural only with *-en* and the plurals that end in *-en* and are not homophones of Dutch verbs. Due to Dutch regular spelling rules, the spelling of the root may differ between the singular and the plural: pluralisation may have resulted in the doubling of the singular's last consonant letter (DOUBLING; e.g., *tak* 'branch' – *takken* 'branches') or removal of the singular's last vowel letter (REMOVAL; e.g., *boot* 'boat' – *boten* 'boats'). Furthermore, both datasets contain plurals that differ from the singular in the voicing of the stem-final obstruent (VOICING; due to final devoicing, e.g., *huis* /fɪœys/ 'house' – *huizen* /fɪœyʒən/ 'houses').

The numbers of word types and the total numbers of responses are provided in Table 1 (first two rows). For Study 1 (DLP), Table 2 provides information about the

**Table 1** Number of stimulus types (and number of responses) per affix and per modality and the overlap in the two modalities

Affix	Modality	Simplex	Complex	Total
<i>-en</i>	Written	932 (30,706)	871 (26,579)	1803 (57,285)
	Spoken	107 (1,804)	143 (2,547)	250 (4,351)
	Written & Spoken	59 (2,885)	86 (4,296)	145 (7,181)
<i>-t</i>	Written	1,140 (3,413)	222 (7,798)	1,362 (11,211)
	Spoken	99 (1,769)	33 (538)	132 (2,307)
	Written & Spoken	9 (469)	7 (391)	16 (860)
<i>-heid</i>	Written	187 (6,514)	55 (1,934)	242 (8,448)
	Spoken	48 (865)	80 (2,480)	128 (3,345)
	Written & Spoken	41 (2,179)	10 (683)	51 (2,862)

**Table 2** Descriptive statistics of the stimuli that were studied in Study 1. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ)

DLP	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	51.70	28.08	54.64	10.58	0.366
SD	135.65	90.83	147.01	42.91	0.258
Min	0.05	0.02	0.02	0.02	0.001
Max	1,617.74	1,470.64	1,617.74	942.69	1.000

root and form frequencies of the stimuli and about the relative frequencies for the plurals. More than half of the responses to plurals were elicited by plurals that do not just consist of the singular plus *-en*: 6,171 responses were given to plurals with DOUBLING (23.22%), 4,905 responses to plurals with REMOVAL (18.45%), and 2,444 responses to plurals with VOICING (9.2%).

For Study 2, Table 3 provides the root and form frequencies of the stimuli and the relative frequencies for the plurals. Less than half of the plurals are not orthographically represented by simply the singular plus *-en*: 579 responses were given to plurals with DOUBLING (22.73%), 687 responses to plurals with REMOVAL (26.97%), and 162 responses to plurals with VOICING (6.36%). Only the voicing is audible in these auditory stimuli.

Studies 1 and 2 are to some extent similar to Baayen, Dijkstra, and Schreuder's (1997b) study, testing the meta model too by investigating how well it can predict RTs. Our subset of DLP includes RTs to 91 of the 93 singular types and 92 of the 93 plural types tested by Baayen, Dijkstra, and Schreuder. These RTs correspond to 6,268 responses in our dataset. The subset of BALDEY includes RTs to only three of the singular types and ten of the plural types that were incorporated by Baayen et al. (1997b). Our datasets are much bigger because we also included plurals that not sim-

**Table 3** Descriptive statistics of the stimuli that were studied in Study 2. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ)

BALDEY	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	26.78	14.90	46.13	12.39	0.346
SD	86.97	77.64	77.64	25.94	0.228
Min	0.14	0.19	0.19	0.10	0.019
Max	574.50	407.17	407.17	181.86	0.955

**Table 4** Descriptive statistics of the stimuli that were analysed in Study 3. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ)

DLP	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	140.02	8.02	151.77	15.33	0.200
SD	254.93	23.91	218.98	30.61	0.174
Min	0.26	0.02	0.52	0.02	0.005
Max	1,337.64	180.50	1,337.64	196.12	0.906

ply consist of the singular plus *-en*, but that also undergo DOUBLING, REMOVAL or DEVOICING.

### 6.1.2 Second/third person singular present tense verb forms ending in *-t*

In Studies 3 and 4, we analysed responses to first person singular present tense verb forms, which just consist of a verb's root (e.g., *bak* 'to bake'; note that in a lexical decision experiment, these verb forms could also be interpreted as imperatives), and second/third person singular present tense verb forms, consisting of the root and *-t* (e.g., *bakt* 'bakes'). We excluded verbs with roots that end in *-t*, because their first, second, and third person singular present tense verb forms are identical. We also excluded the verbs *hebben* 'to have', *kunnen* 'can', *worden* 'to be (passive voice)', and *zullen* 'will', because they can function as auxiliary verbs and therefore have exceptionally high root frequencies. Table 1 shows the numbers of word types and total numbers of responses analysed in Study 3 and 4 (third and fourth row). The root and form frequencies of the stimuli and the relative frequencies for the plurals are provided in Table 4 for Study 3 and in Table 5 for Study 4.

### 6.1.3 Derived nouns ending in *-heid*

In Studies 5 and 6, we analysed responses to uninflected adjectives that can combine with *-heid* and nouns that consist of an adjective root and this suffix. Table 1 (fifth and last row) provides the numbers of word types and responses that were analysed.

**Table 5** Descriptive statistics of the stimuli that were analysed in Study 4. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ)

BALDEY	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	123.21	21.21	35.21	5.17	0.117
SD	371.48	74.51	84.01	16.79	0.131
Min	0.29	0.05	0.55	0.05	0.005
Max	2,484.93	559.88	477.69	91.62	0.605

**Table 6** Descriptive statistics of the stimuli that were analysed in Study 5. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ)

DLP	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	101.77	46.63	205.03	8.43	0.113
SD	223.06	114.86	363.72	18.46	0.139
Min	0.81	0.1	4.98	1.02	0.002
Max	1,835.81	1,091.10	1,835.81	120.71	0.675

**Table 7** Descriptive statistics of the stimuli that were studied in Study 6. The relative frequency of complex words (Rel. FQ) refers to the complex words' form frequencies (Form FQ) divided by the complex words' root frequencies (Root FQ)

BALDEY	Simplex words		Complex words		
	Root FQ	Form FQ	Root FQ	Form FQ	Rel. FQ
Mean	74.33	44.61	91.33	2.62	0.258
SD	185.76	119.31	258.52	4.59	0.358
Min	1.17	0.67	0.05	0.05	0.001
Max	1206.36	751.40	1835.81	33.14	1.000

Table 6 provides frequency information about the stimuli analysed in Study 5 and Table 7 about the stimuli analysed in Study 6.

## 6.2 Model fitting

All statistical models described in this paper were implemented in R (R Core Team, 2021). We implemented formulae (1), (2), (3), representing the three models that we test, as Bayesian models, written in the language *stan* (Stan Development Team, 2022). We fitted the models with the package *cmdstanR* (e.g., Gabry & Češnovar, 2022) and its default Hamiltonian Monte Carlo (HMC) algorithm. We used four chains and as many samples as were needed to ensure that all values of R-hat were

less than 1.01 and all effective sample sizes were above 400, which safeguards that the models' estimates are stable (using the criteria in Vehtari et al., 2021).

### 6.2.1 Model comparison

We compared the models' performances to see which one best accounted for the data. Model comparisons were based on the models' expected log predictive density (ELPD LOO; Vehtari et al., 2017), which provides a measure of predictive performance on unseen data. The higher the ELPD LOO, the better the model generalises to unseen data. We will graphically present how much the best performing model differs in ELPD LOO from the other models, accompanied by the standard error (SE) of these differences. Traditionally, these differences are expressed in the form of the ELPD LOO of a model minus the ELPD LOO of the best model (rather than the other way around; e.g., Gravel et al., 2024). Thus, differences in ELPD LOO are depicted as negative values.

Vehtari et al. (2017) favoured a model over another in case of a ELPD LOO difference of  $-10.2$ , which was twice as large as this difference's SE of 5.1, but they did not give any explanation for this threshold. We follow their criterion and assume that two models only systematically differ in their predictive performance when the absolute difference in ELPD LOO is at least twice the corresponding SE. In addition, because small differences in ELPD LOO are less informative (Sivula et al., 2020), we assume that the performance of two models can only be statistically distinguished if the difference in ELPD LOO is greater than  $-4$  (cf. Vehtari, 2020). Because the ELPD LOO does not penalise model complexity (Gronau & Wagenmakers, 2019), we prefer the less complex model when two models yield similar predictive performance according to our ELPD LOO criterion.

The data set of Study 1 is too large for the ELPD LOOs to be computed. In this case, we compute the ELPD LOO on the basis of a random subsample of the data that consists of one tenth of all observations, following Magnusson et al. (2020; see also Appendix B).

### 6.2.2 Prior estimation

Fitting Bayesian models requires the specification of prior distributions (henceforth *priors*) for the models' parameters (e.g., the effects of the frequency measures on the RTs and the size of the parsing penalty in the tipping point model and the root-driven model). These priors represent assumptions about possible and likely values of these parameters.

We determined the priors for the effects of length in number of letters / ms, trial number, form frequency, and root frequency as well as by-participants random intercepts on the basis of the RTs of the correct responses to all morphologically simplex and complex stimuli in DLP and BALDEY, respectively, that closely resemble but are not identical to the stimuli to be studied in Studies 1-6. We analysed the RTs to these responses with Bayesian linear mixed-effects models, determining the coefficients of the predictors for which we wished to determine the priors. We chose uninformative priors for the predictors, that is, normal distributions with mean = 0, SD = 0.2, and

**Table 8** Estimated effect sizes and standard deviations (SD) for the variables of interest in DLP. These estimates were used as priors in Studies 1, 3, and 5

	Intercept	Length	Trial	Root frequency	Form frequency
Coefficient	6.5173	-0.0075	0.0075	-0.0042	-0.0213
Coefficient SD	0.0658	0.0026	0.0006	0.0003	0.0003

**Table 9** Estimated effects and standard deviations (SD) for the variables of interest in BALDEY. These estimates were used as priors for models applied to BALDEY in Study 2, 4, 6

	Intercept	Duration	Trial	Root frequency	Form frequency
Coefficient	5.7090	0.2176	0.0127	-0.0054	-0.0091
Coefficient SD	0.1016	0.0125	0.0029	0.0018	0.0016

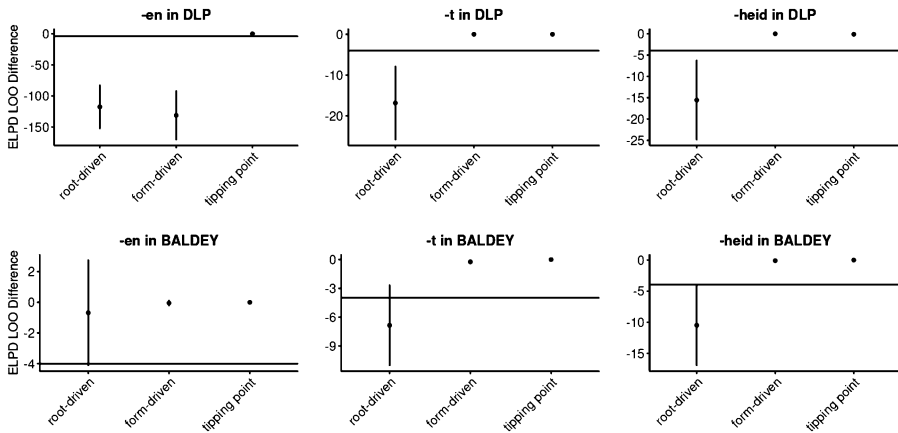
a positive prior of  $10^{0.1}$  for the SD of the by-participant intercept. As in Studies 1-6, we excluded RTs that were smaller than 100 ms from word onset (DLP) or from word offset (BALDEY), and log-transformed all numerical variables.

The dataset from the DLP contains 160,906 responses with 116,320 responses to 3,627 unique simplex words and 44,586 responses to 1,436 unique suffixed words. Table 8 shows that word length in letters, root frequency, and form frequency yield facilitative effects, whereas the effect of trial number is inhibitory. The estimated coefficients shown in Table 8 were used as priors for all models reported below that were fitted to RTs in DLP.

The dataset for determining the priors for BALDEY contains 7,556 responses, with 378 responses to 21 unique simplex words and 7,556 responses to 440 unique suffixed words. Table 9 shows that the estimated effects of duration in ms and trial number are inhibitory, while root and form frequency yield facilitative effects. The estimated effects shown in Table 9 were used as priors for all models reported below that were fitted to RTs of BALDEY.

We tested different priors for the parsing penalty, because we had no knowledge about nor strong indications of what would be good priors for the parsing penalty and we wished to rule out that the priors for the parsing penalty affect the estimate of the parsing penalty, and consequently the goodness of the model's fit to the data. Assuming that recognising a complex word based on its root (and morphologically related words) takes between 1 ms and 350 ms, we tested 350 normal distributions as parsing penalty priors, each with a mean ranging between 1 ms and 350 ms and a standard deviation  $SD_{prior}$  as given in Equation (4). In Equation (4), both mean and SD are expressed in ms. Equation (4) ensures the standard deviation is approximately one fourth of the mean for larger means, but is never shorter than 1 ms. If SDs were smaller than 1 ms, the prior distributions would be too peaked so that the information in the data could not overrule the information provided by the prior. For more details about how we computed the priors, see Appendix C.

$$SD_{prior} = mean/4 + 4/(mean + 4) \quad (4)$$



**Fig. 1** Difference in ELPD LOO (y-axis) between the three processing models (x-axis) and the model(s) with the highest ELPD LOO (which can be two models at the same time), for each suffix (columns) in the two modalities (rows: visual in top row, auditory in bottom row). Error bars represent two times the standard error. The solid horizontal line in each panel represents a difference in ELPD LOO of minus four, our threshold for determining whether two models significantly differ in how accurately they fit the data

## 7 Results

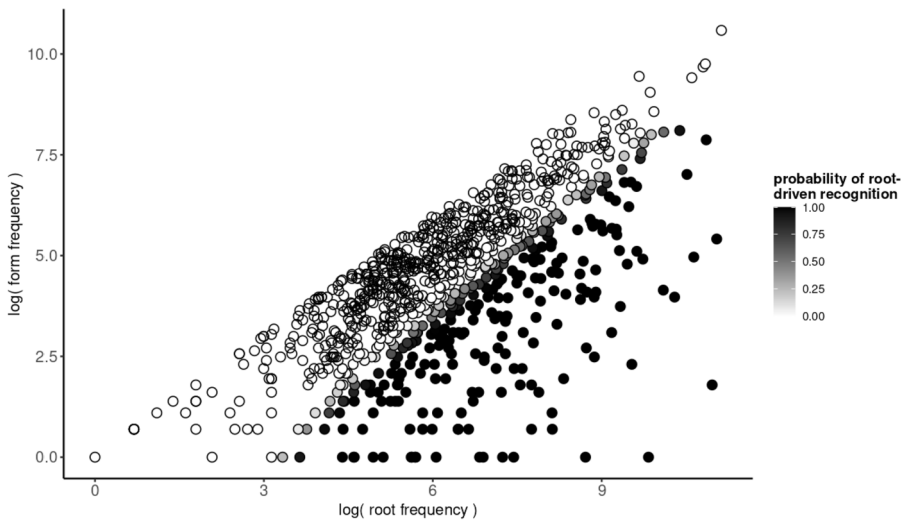
Figure 1 shows, for Studies 1-6, the difference in ELPD LOO between the root-driven model, the form-driven model, and the tipping point model relative to the model that predicts the RTs the most accurately (i.e., the model with the highest ELPD LOO). In Study 1 (written *-en* in DLP), the tipping point model predicts the RTs most accurately. In Study 2 (spoken *-en* in BALDEY), the root-driven and the form-driven model predict the RTs equally accurately; the tipping point model may yield the same predictive accuracy as the other two models, but it is more complex and thus considered to be a worse fit to the data.

In Studies 3-6 (written/spoken *-t/-heid* in DLP/BALDEY), the form-driven model predicts the RTs most accurately. The tipping point model is considered to fit the data worse than the form-driven model in Studies 3-6 because it draws on a greater complexity to yield the same prediction accuracy as the form-driven model. Moreover, the tipping point model actually behaves like the form-driven model because the parsing penalty is so high that the recognition of complex words is always form-driven and never root-driven (see Appendix B).

The results from all studies are independent of the choice of the prior mean for the parsing penalty. We describe the relevant aspects of the results of the six studies in more detail in the following subsections. Summaries of all models fitted can be found in Appendix D. The coefficients of the control variables and the reliability of their estimations are not further discussed here as they are not of interest for answering our research questions.

### 7.1 Study 1: written plural nouns ending in *-en*

As mentioned above (and as also shown in Fig. 1), RTs to written nouns ending in *-en* are predicted the most accurately by the tipping point model. In this model, the



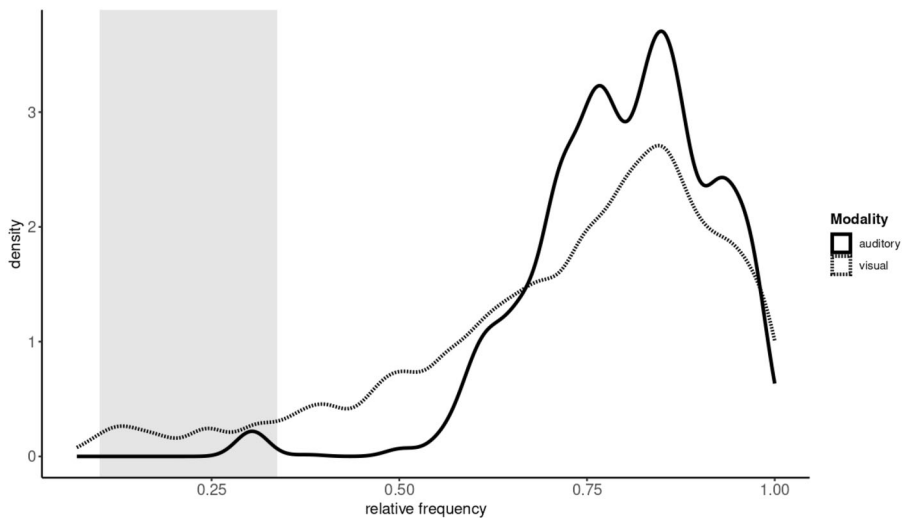
**Fig. 2** Probabilities of root-driven recognition (indicated by marker grey level) for written plural nouns ending in *-en* as a function of the plurals' log-transformed root frequencies (x-axis) and form frequencies (y-axis). The probabilities are based on the posterior distribution of the tipping point model. Clearly, the tipping point depends on both root frequency and form frequency

estimated effects of both root frequency (mean =  $-0.0291$ , SD = 0.0005) and form frequency (mean =  $-0.0261$ , SD = 0.0006) are facilitative. The estimated parsing penalty for the tipping point model is 59 ms. With this estimated parsing penalty, the recognition of 14% of the plurals is always root-driven and the recognition of 28% of the plurals is always form-driven. For 58% of the plurals, both root-driven and form-driven recognition is possible.

As can be seen in Fig. 2, the root-driven recognition only occurs when the complex word's log-transformed root frequency is at least 3.14 (i.e., 0.55 per million tokens). Because of the meta model's mathematical formulation, whether recognition is root-driven or form-driven depends on the relative frequency (i.e., form frequency divided by root frequency). Our results suggest that relative frequencies lower than 0.101 result in obligatory root-driven recognition and relative frequencies higher than 0.337 result in obligatory form-driven recognition. Plural nouns with relative frequencies between these two values can be recognised with either processing mechanism.

## 7.2 Study 2: spoken plural nouns ending in *-en*

As mentioned above (and see also Fig. 1), RTs to spoken nouns forming their plural with *-en* are equally accurately predicted by the root-driven model and the form-driven model (the tipping point model relies on a greater complexity than the other two models and is thus disregarded). The estimated effect of root frequency in the root-driven model (mean =  $-0.0063$ , SD = 0.0012) is facilitative, as is the estimated effect of form frequency in the form-driven model (mean =  $-0.0053$ , SD = 0.0014). For plurals, log-transformed root and form frequency strongly correlate with each other ( $r(2,545) = .864$ ,  $p < .001$ ), and, therefore, one may argue that the similar



**Fig. 3** Plurals' relative frequencies (x-axis) and corresponding densities (y-axis) for DLP and BALDEY (line type). The grey-shaded area marks relative frequencies between 0.101 and 0.337, which were associated with flexible processing mechanisms (root- and form-driven) in Study 1

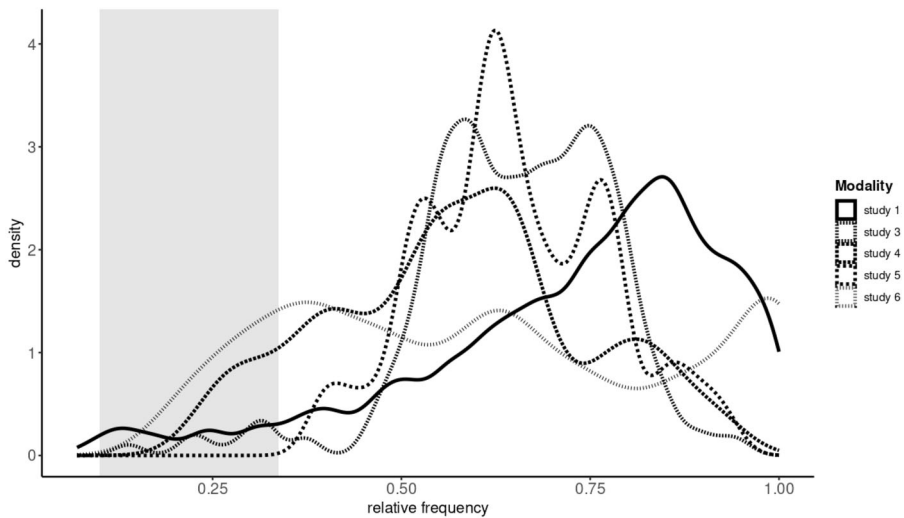
performance of the root-driven model and the form-driven model may be due to a high correlation between root and form frequency. Interestingly, a similarly strong correlation was present in the materials of Study 1 ( $r(57,283) = .744, p < .001$ ) and, in that study, the root- and form-driven models also performed equally well.

Figure 3 shows the relative frequencies of the plural nouns in the written (Study 1) and spoken (Study 2) datasets. Compared to the visual dataset, the auditory dataset provides poorer coverage of verb forms with relative frequencies between 0.101 and 0.337 – forms that, as discussed in Sect. 7.1, can be processed either root- or form-driven. This may be the reason why the tipping point model is supported by the visual data but not by the auditory data.

### 7.3 Studies 3-6: written/spoken words ending in *-t*/*-heid*

As mentioned above (and shown in Fig. 1), RTs to second/third person singular present tense verb forms ending in *-t* and RTs to derived nouns ending in *-heid* are the most accurately predicted by the form-driven model in both the visual and the auditory modality. The effect of form frequency is facilitative regardless of suffix and modality (see Tables 12-17 in Appendix D). Correspondingly, Studies 3-6 do not provide support for the tipping point model.

Figure 4 visualises the density distributions of the relative frequencies of the complex words examined in Studies 3-6, with Study 1 included as reference. It shows that stimuli with relative frequencies between 0.101 and 0.337 – those that can be processed via either their root or form (cf. Sect. 7.1) – make up a smaller proportion of the stimulus set in Studies 3 and 5 than in Study 1. In contrast, this frequency range is more heavily represented in Studies 4 and 6. Crucially, the fact that the tipping point model does not always account the best for the data, even when many



**Fig. 4** Complex words' relative frequencies (x-axis) and corresponding densities (y-axis) for Studies 1 and 3–6 (line type). The grey-shaded area marks relative frequencies between 0.101 and 0.337, which were associated with flexible processing mechanisms (root- and form-driven) in Study 1

stimuli fall within this range, suggests that relative frequency alone does not fully determine model fit.

## 8 General discussion

We investigated how root frequency and form frequency of three types of Dutch morphologically complex words affect how they are recognised. Specifically, we examined whether there is a tipping point in the relative frequency of root versus form that determines whether processing is predominantly root-driven or form-driven. To this end, we analysed lexical decision reaction times (RTs) to both written and spoken words consisting of a root and one of three suffixes (*-en*, *-t*, *-heid*). These RTs were modelled using three different statistical approaches: one reflecting the tipping point hypothesis, and two others representing the assumptions that processing is either consistently root-driven or consistently form-driven.

We found that four of the six analysed datasets (written and spoken *-t/-heid*) are best fitted with a form-driven account. One dataset (spoken *-en*), for which form and root frequency are highly correlated, is equally well predicted by both the root-driven and the form-driven account. These results lend support to the idea that surface forms play a more important role in the word recognition process than morphological roots.

Interestingly, visually presented nouns ending in *-en* lend support to the tipping point hypothesis. Word recognition is consistently form-driven for plurals with root frequencies that are less than 3 times higher than their form frequencies, and root-driven for words with root frequencies that are more than 10 times higher than the corresponding form frequencies. The recognition of visually presented plurals with

root frequencies that are about 3 to 10 times higher than their form frequencies (relative frequency between 0.101 and 0.337) can be both root- and form-driven. This finding suggests that both morphological roots and surface form may be of importance for the word recognition process.

Our finding that written Dutch plural nouns ending in *-en* support both root- and form-driven processing dovetails with the results reported by Baayen et al. (1997b). This raises the question of why plural nouns ending in *-en* lend support to both root-driven and form-driven processing, whereas second/third person singular present tense forms ending in the suffix *-t*, and nouns that are derived from adjectives with the derivational suffix *-heid* support form-driven processing in both the visual and the auditory modality.

### 8.1 The exceptional status of nouns ending in the plural suffix *-en*

Both the plurals ending in *-en* and the verb forms ending in *-t* are inflectional forms and/or semantically transparent (in contrast to *-heid*, which is derivational and semantically less transparent). This rules out that one of these properties is the reason that plurals are recognised differently. This also holds when, in accordance with Booij (1995), *-en* is considered as an instance of inherent inflection, while *-t* is an instance of contextual inflection. Because inherent inflection is generally considered to be closer to derivation than contextual inflection, one might then predict, if anything, that *-en* patterns with *-heid* and *-t* to stand out, which is not the case.

The affixes *-en* and *-heid* both lead to resyllabification of the word and both contain a vowel. So resyllabification and the presence of a vowel also do not seem to account for the mode of processing. Baayen et al. (1997b) suggest that the lower a complex word's relative frequency is, that is, the lower the complex word's frequency is in comparison to its root frequency, the more likely the complex word undergoes decomposition (cf. also Hay, 2001; Hay & Baayen, 2002). As can be seen in Table 3, 8, and 11, the *-en* stimuli have higher relative frequencies (mean = 0.366, SD = 0.258) compared to both the *-t* stimuli (mean = 0.2, SD = 0.174) and *-heid* stimuli (mean = 0.113, SD = 0.258). Our results therefore also do not support the hypothesis that a low relative frequency may lead to a role for root frequency in the recognition of at least some forms.

It is especially the plurals with relative frequencies between 0.101 and 0.337 that show effects of both root and form frequencies. Plurals with lower relative frequencies mostly show root frequency effects and plurals with higher relative frequencies mostly form frequency effects. The visual plurals in the relative frequency range from 0.101 to 0.337 do not seem to have other properties than the other plurals, except for relative frequency. The vast majority of the spoken plurals and the *-t* and *-heid* words presented visually have relative frequencies that are (much) higher than 0.337. This could explain why these words do not show root frequency effects. The orally presented *-t* and *-heid*, in contrast, are represented with many words with frequencies below 0.337. Possibly, for spoken words, root frequency plays a less important role than for written words, because the root is less easy to segment from spoken words, as a result of the acoustic differences between roots spoken in isolation and in different complex words (including differences in duration and co-articulation) that are not visible in the spellings of these Dutch words.

Another explanation for why written *-en* favours the tipping point hypothesis in contrast to *-t* and *-heid* is that *-en* is more productive than *-t* and *-heid*. It does not only form plurals for nouns, but also for verbs, and it can be used to derive nouns from verbs. A relationship between productivity, that is, how frequent an affix is and how diverse an affix' contexts of use are, and mode of processing forms another explanation of the finding that the form-driven and the root-driven model predict RTs to spoken words ending in *-en* equally accurately. There are indications that affix productivity increases the likelihood of root-driven processing in Dutch and English (Bertram et al., 2000; Bertram et al., 1999; Hay & Baayen, 2002). In Arabic, which relies heavily on non-affixal morphology, the productivity of root templates appears to facilitate root-driven processing as well (Boudelaa & Marslen-Wilson, 2011), pointing again to a relationship between productivity and processing mode.

## 8.2 The precise formulation of the tipping point model

Our tipping point model, based on the meta model by Baayen et al. (1997b), predicts that the roles of root and form frequencies mostly depend on the basis of the relative values of these frequencies. The question arises whether this formulation of the tipping point model is too simplistic to predict when root and form frequencies contribute to a word's recognition. Our results suggest that there is a difference between the visual and the auditory modality. In addition, there may be a role for the productivity of the affix. A new and more elaborated formulation of the tipping point model may take these factors into account.

Another issue with our implementation of the tipping point model concerns the exact definition of the root frequency. In the present study, we determined root frequency on the basis of words consisting of at most two morphemes. Including also words that contain more than two morphemes yields comparable frequencies as indicated by high correlation coefficients between the two frequency measures ( $0.94 < r < 0.99$ , depending on the analysed dataset), because of which it may not make a difference whether one includes words consisting of more than two morphemes or not. Such a difference may be observed, though, in the context of highly agglutinative languages, in which words regularly consist of more than two morphemes.

Although the tipping point model only performs better than a root-driven or a form-driven model for one of the analysed datasets, this is a clear indication of a trade-off between root-driven and form-driven processing. In light of our results for the visually presented plurals which are sometimes recognised by their roots and sometimes by their forms, it is questionable though, whether such a trade-off is categorical in nature (i.e., recognition is either root-driven or form-driven), as assumed in our tipping point model. It could be that root-driven and form-driven processing are two poles of a continuum with mixed processing mechanisms in between.

## 9 Conclusion

In the present study, we tested for the first time on the same datasets whether a tipping point model better describes the processing of both written and spoken morphologically complex words than a root-driven and a form-driven model. We formulated

the three theoretical processing models as statistical models following Baayen et al. (1997b) and identified the models that best fit the distributions of human lexical decision data for Dutch complex words. Our results suggest that the recognition of most complex words, in both visual and auditory word recognition, is form-driven, but root-driven processes seem to play a role too, at least in the processing of visually presented plural nouns ending in *-en*. We argue that relative frequency does not determine on its own whether recognition is root- or form-driven, but that it interacts with other factors such as the ease with which the root can be identified and affix productivity. Moreover, our results do not support a categorical tipping point that determines whether word recognition is either root-driven or form-driven, but rather a continuum between root- and form-driven recognition. Because distributional-connectionist theories do not posit a strict boundary between processing modes, whereas localist theories often do, accounting for such a continuum seems to be a bigger challenge for localist theories than distributional-connectionist theories of human word recognition.

## Appendix A

We assume that the predicted RT to a simplex word is the sum of the intercept  $\beta_0$ , a set of predictors  $x_1, \dots, x_i$  for  $I$  predictors, multiplied with their corresponding betas  $\beta_1, \dots, \beta_i$ , and a by-participant intercept  $u_j$  for  $J$  participants. A word's root and form frequency belong to the set of predictors  $x_1, \dots, x_i$ . Simplex words are always recognised through their roots, which are identical to their whole forms, regardless of the processing model. Because RTs to simplex words depend on the root frequency  $FQ_{root}$  (Baayen et al., 1997b), the RT to a simplex word is predicted by

$$RT_{Simplex} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \beta_{FQ_{root}} FQ_{root} + u_j. \quad (1)$$

The root-driven, form-driven, and tipping point model produce different predictions for a complex word's RT. In the root-driven model, recognition of complex words is always root-driven, which implies that plural recognition involves the root frequency plus an additional parsing penalty  $\Delta p$ , resulting in the following predicted response time:

$$RT_{Complex} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \beta_{FQ_{root}} FQ_{root} + \Delta p + u_j, \quad (2)$$

The form-driven model, in contrast, assumes that recognition of complex words is always form-driven and that the time needed for recognising the whole form is a function of the form frequency  $FQ_{form}$ , without parsing penalty. Thus, the predicted RT to a complex word reads:

$$RT_{Complex} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \beta_{FQ_{form}} FQ_{form} + u_j. \quad (3)$$

The tipping point model combines formulae (2)–(3). Depending on which processing mode would predict the quickest response, the equation for predicting the RTs to complex words is based on the form frequency (if form-driven recognition is faster)

or the root frequency plus an additional parsing penalty (if root-driven processing is faster):

$$RT_{Complex} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \min(\beta_{FQ_{form}} FQ_{form}, \beta_{FQ_{root}} FQ_{root} + \Delta p) + u_j. \quad (4)$$

During model fitting,  $x_1, \dots, x_i$  are provided so that  $\beta_0, \dots, \beta_i, u_j$ , and  $\Delta p$  can be estimated. This estimation is comparable to fitting a linear mixed-effects model (cf. Sorensen & Vasishth, 2015), except for the *min*-argument, which is necessary to determine the fastest processing mode.

## Appendix B

To prove that our methodological approach can correctly determine whether a distribution of RTs has likely been generated by a root-driven model, a form-driven model, or a tipping point model, we conducted a dedicated validation study. For doing so, we used a two-step procedure. First, we synthesised RTs with three theoretical processing models based on each of the three processing views (i.e., with formulae (1)–(3)). Second, we investigated whether the theoretical processing model used for synthesising the RTs also most accurately predicts these RTs.

### B.1 Data synthesis

We generated RTs from invented, henceforth synthesised, participants reacting to invented, henceforth synthesised, simplex and complex words. For doing so, we assigned properties to the synthesised participants and synthesised words that play a role in the formulae of the three theoretical processing models, that is, root frequency, form frequency and word length/duration (see (1)–(3)). We wanted the generated RTs' distributions to closely resemble the RTs' distributions of simplex and suffixed words that are to be analysed in the subsequent studies, which helps to ensure that the conclusions drawn are valid. To ensure the greatest comparability, we based these properties of our synthesised words (i.e., words' lengths/durations, root frequencies, and form frequencies;  $x_1$ ,  $F_{root}$ , and  $F_{form}$  in the formulae) on nouns in BALDEY (Ernestus & Cutler, 2015). In addition, we determined, also on the basis of DLP and BALDEY, what the effects of these properties were on the participants' RTs to the synthesised words. We will now discuss each of these steps in detail.

We generated 100 sets of RTs with each of the three models (1-3). Each set consists of 6,000 RTs, produced by 20 synthesised participants, responding to 150 synthesised simplex words and 150 synthesised complex words. We based the properties of our synthesised words (i.e., words' durations, root frequencies, and form frequencies;  $x_1$ ,  $F_{root}$ , and  $F_{form}$  in the formulae) on nouns in BALDEY (Ernestus & Cutler, 2015). We only analysed the 504 nouns that can form their plural only according to the scheme *root* + *-s* and that cannot occur as verbs. Table 10 lists descriptive statistics of the words' form frequencies, root frequencies and durations.

In order to determine the effect sizes of the variables in our generating models (1-3), that is, of duration, root frequency, form frequency, trial number (i.e.,  $\beta_1, \beta_2, F_{root}$ ,

**Table 10** Means and standard deviations (SD) of the words' durations, root frequencies, and form frequencies, on which we based the properties of our synthesised words in the simulation experiment

	Mean	SD
Form frequency	4.118611	1.7497953
Root frequency	5.841478	1.9881844
Duration	6.485120	0.2692574

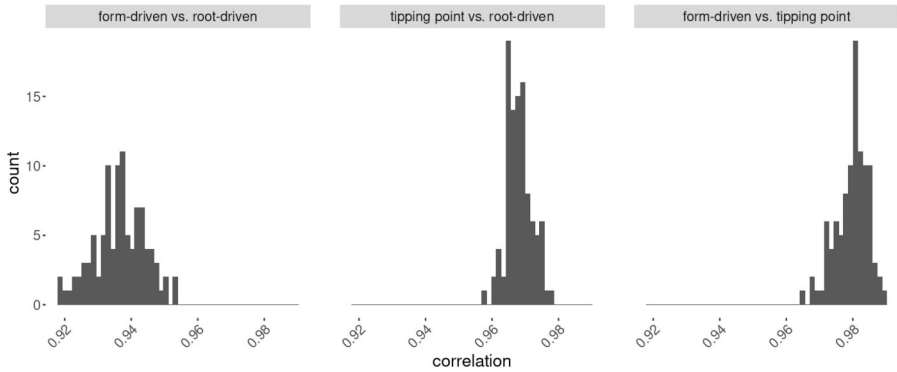
**Table 11** Estimated effect sizes with corresponding standard deviations (SD) for the effects in BALDEY that are incorporated in the synthesis of RTs in the simulation experiment

	Coefficient	Coefficient SD
Intercept	5.620486325	0.089528972
By-participant effect	0.000000000	0.242652274
Duration	0.236941020	0.010009395
Trial	0.009320023	0.002562212
Root frequency	-0.002597006	0.001428608
Form frequency	-0.017916674	0.001626829

and  $F_{form}$ ), and by-participant intercepts ( $u_j$ ), we determined their effect sizes on the RTs for the nouns in BALDEY on the basis of which we had also determined the frequencies and durations of our synthesised words (as listed in Table 10). We excluded incorrect responses, responses made before word offset, and seven observations made in session 8 by Participant 1 because of erroneous encodings. The resulting constrained dataset comprises 9,742 observations. We fitted a Bayesian linear mixed-effects model to the log-transformed RTs, including form frequency, root frequency, trial number and duration, as predictors, all of which were also log-transformed, as well as by-participant random intercepts. The resulting estimates are presented in Table 11.

We then generated the RTs of a single set as follows. We assumed 300 words with characteristics randomly sampled from the distributions shown in Table 10 and ascribed a unique trial number in the interval  $\{\log(1), \log(2), \dots, \log(300)\}$ . For instance, a synthesised word could have a duration of  $e^6$ , which would correspond to 403 ms. For determining how much the synthesised duration contributes to the synthesised RT, the duration would be multiplied with 0.237, which is the effect of duration (see Table 11). Thus, 96 ms of the synthesised RT are due to the synthesised word's duration. Similarly, we randomly sampled 20 by-participant random intercepts from a normal distribution with mean = 0 and the estimated SD reported in Table 11 (0.243). Using the effect sizes' estimates listed in Table 11, we thus generated 300 RTs for each of the 20 synthesised participants with each of the three theoretical processing models (formulae (1)–(3)). We repeated this procedure 100 times, resulting in  $100 \times 6,000 = 600,000$  generated RTs for each model.

Because the estimated root frequency effect is relatively small in comparison to the estimated form frequency effect (Table 11), our generated RTs to the assumed complex words would never be produced via the root-driven mechanism in our tip-



**Fig. 5** Histogram of the coefficients of the correlations (x-axis) between the 100 pairs of synthesised response times that were synthesised with the root-driven and the form-driven model (first column), with the root-driven and the tipping point model (second column), and with the form-driven and the tipping point model (third column)

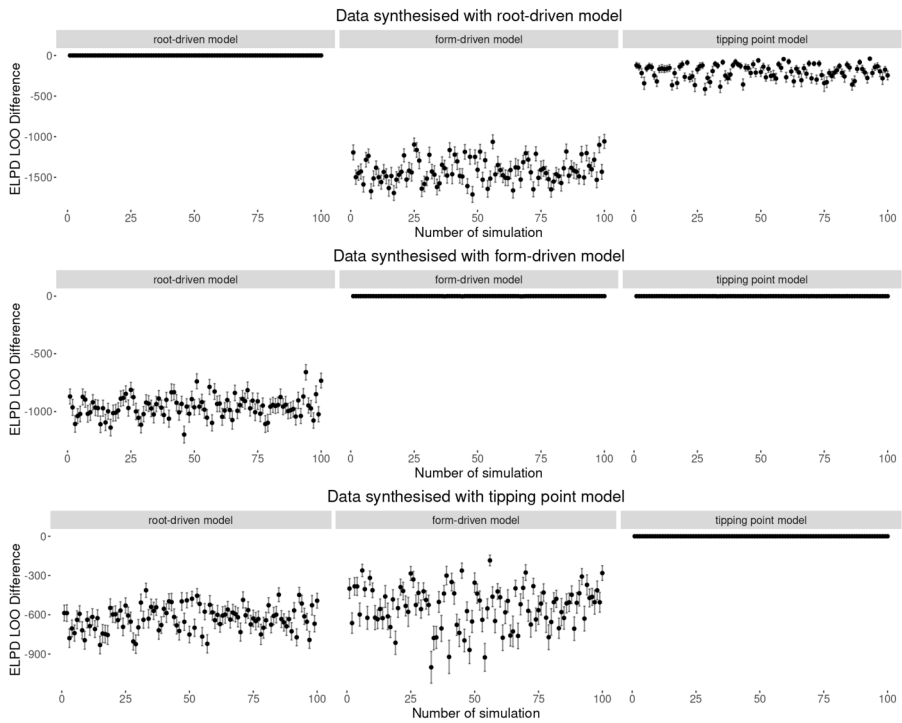
ping point model. To ensure that the recognition of at least some RTs to complex words were root-driven, we multiplied the effect estimate for root frequency with 50. As a consequence, the tipping point model produced complex words via root-driven mechanisms across the 100 repetitions of the generating procedure between 30% and 49% of the trials.

## B.2 Analysis of synthesised data

We fitted each theoretical processing model to the 300 sets of generated RTs to determine its goodness of fit with each set. We provided the model with the generated words' variables and their RTs and let the model estimate the variables' effect sizes and the (by-participant) intercept(s). We used as priors for our models the distributions of the effects that were used for generating the RTs (and which were deduced from the actual data).

## B.3 Results

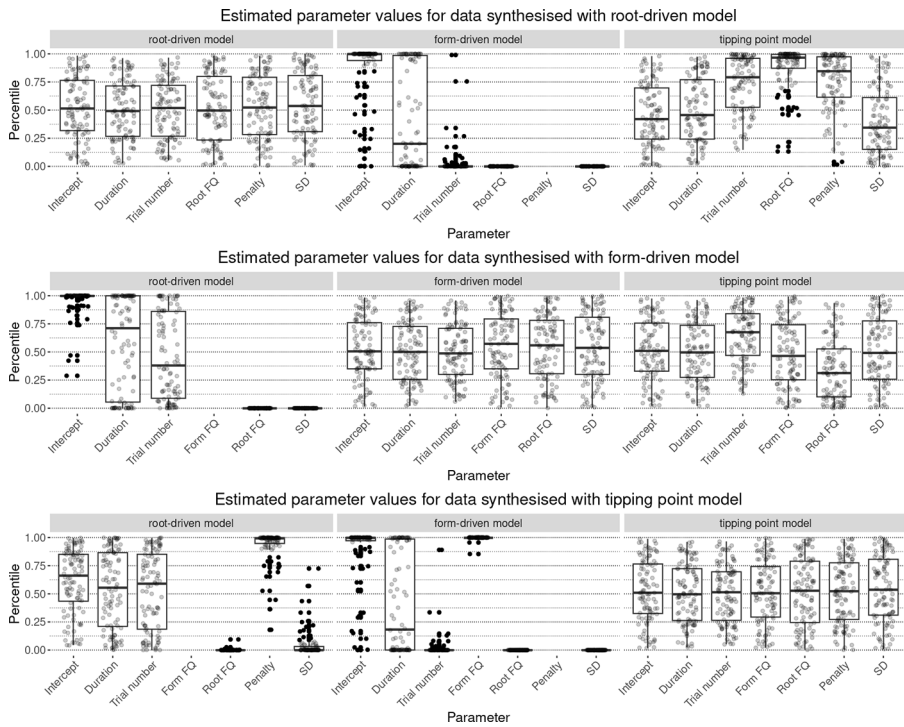
We first investigated whether the sets of RTs generated with the three different models are substantially different. Only if they are different, we can expect that it is possible to determine by which model a set was generated. For each pair of models (root-driven versus form-driven, root-driven versus tipping point, and form-driven versus tipping point), we compared the correlation between each of the 100 RT sets generated by the first model with the corresponding set generated by the other model, which resulted in 100 correlation coefficients for each pair of models. Figure 5 is divided into three cells, each showing a histogram that represents the 100 correlation coefficients for one pair of models. For example, the histogram in the leftmost cell shows correlation coefficients between the 100 pairs of RT sets generated with the root-driven (100 sets) and the tipping point model (100 sets). All correlation coefficients range between  $r_{min} = .92$  and  $r_{max} = .99$  and the correlations between RT



**Fig. 6** Difference in ELPD LOO (y-axis) between the root-driven model (first column), the form-driven model (second column), and the tipping point model (third column) relative to the model with the highest ELPD LOO, for each synthesised distribution of 6,000 response times (x-axis), when the data was synthesised by the root-driven model (top row), the form-driven model (second row), and the tipping point model (bottom row). Error bars represent two times the standard error. The figure shows that the model that synthesised the data also fits the data the most accurately

distributions from the form-driven model and the tipping point model are especially high. The overall strong correlations indicate that the models generate relatively similar RTs, which makes it difficult to tease the generating models apart on the basis of the RT distributions. The RT distributions are so similar because the RTs are mostly determined by the control predictors (e.g., a word's duration) and less so by the processing mode of a complex word (root-driven or form-driven). The RTs from the form-driven and the tipping point model are particularly similar because the tipping point model, when incorporating a relatively high parsing penalty, can behave like a form-driven model.

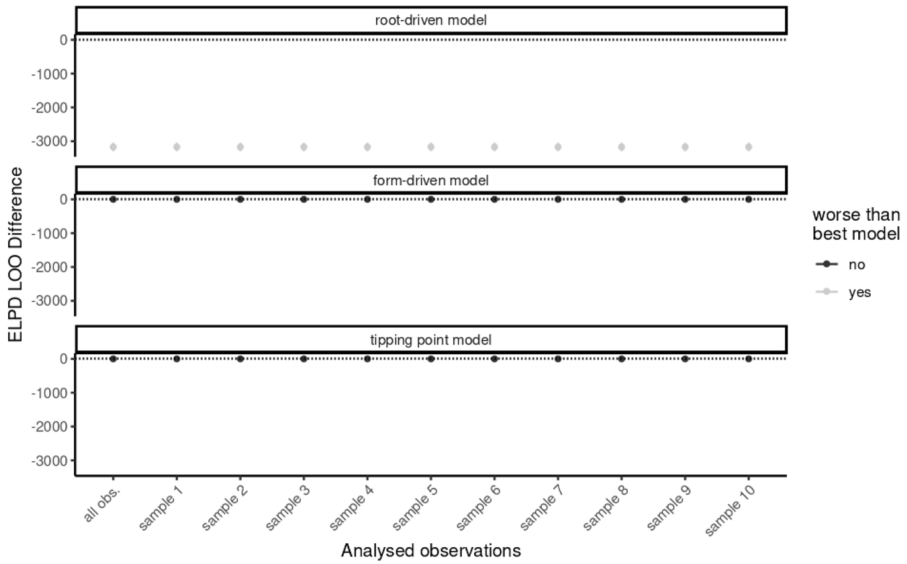
We then investigated whether the original generating models can be determined on the basis of the RT distributions. Figure 6 shows the difference in ELPD LOO between the root-driven model, the form-driven model, and the tipping point model relative to the model with the highest ELPD LOO, for each set of generated RT distributions. The figure is split up into nine cells, which cross the type of generating model (in rows) and the type of inferential model tested (in columns). Each value in a cell on the x-axis represents one of the 100 generated distributions of RTs produced by the pertinent model. It can be seen that, although the RT distributions from the



**Fig. 7** Boxplots of the percentiles of the known effect sizes (y-axis) in the posterior distributions for these effect sizes (x-axis) as estimated by the root-driven model (left column), the form-driven model (second column), and the tipping point model (third column) when the data was synthesised by the root-driven model (top row), form-driven model (second row), and tipping point model (bottom row). Each dot represents the results from the synthesised distributions consisting of 6,000 trials

three models are highly correlated, the model that generated the RTs makes the most accurate predictions of the RT distributions. The exception is formed by the RTs generated by the form-driven model, which are as accurately fitted by the tipping point model as by the form-driven model. This is because, as mentioned above, the tipping point model can effectively function as a form-driven model when the parsing penalty is so high that the recognition of plurals is always form-driven. In those cases, we prefer the form-driven model, because it is less complex than the tipping point model. Given these results, we conclude that, with our approach, it is possible to correctly infer the model that generated the RT distribution.

Some of the datasets that we analyse in the subsequent studies contain fewer data points than the datasets tested in Study 1. It is therefore important to note that we established that the same conclusions are reached when Study 1 is carried out with smaller sets, consisting of 2,000 RTs (instead of 6,000 RTs), produced by 20 synthesised participants, responding to 100 synthesised words (instead of 300 synthesised words). That is, we found again that the model that generated the RTs makes the most accurate predictions of the RT distributions (Fig. 7). The exceptions are, again, the RTs generated by the form-driven model, which are predicted as accurately by the tipping point model as by the form-driven model.



**Fig. 8** Difference in ELPD LOO (y-axis) between the root-driven model (first row), the form-driven model (second row), and the tipping point model (third row) relative to the model with the highest ELPD LOO (which is for each mean parsing penalty the form-driven model and the tipping point model), for all observations of a given dataset (x-axis, very first column) and ten random subsamples (x-axis, other columns with labels *sample 1*, *sample 2*, ..., *sample 10*). The two horizontal dotted lines in each panel represent differences in ELPD LOO of zero and minus four. Error bars represent two times the standard error. The figure shows that the difference in ELPD LOO between the three models does not change when based on a subset of the data

One of the datasets that we analyse in Study 1 contains so many observations that it is infeasible for us to include all the data in the analysis, because the computation of the ELPD LOOs for this dataset exceeds the working memory capacity of 512 GB of our largest computing node. Following Magnusson et al. (2020), we compute the ELPD LOO on the basis of a random subsample of the data that consists of one tenth of all observations. To show that model comparisons are stable regardless of whether the ELPD LOO is based on the whole dataset or on a subset of 10%, we computed the ELPD LOO for the first dataset that was generated with the form-driven model both on the basis of all data points in that set and on the basis of ten different subsamples, each consisting of one tenth of the whole dataset. Figure 8 shows that the difference in ELPD LOO between the three inferential models is approximately the same regardless of whether the ELPD LOO is based on all observations (first column) or on one of ten different subsets (other columns).

In conclusion, the simulation experiment shows that, although the RT distributions generated by the three models are highly correlated, it is possible to infer the generating model, in the controlled environment of generated RTs. Because we generated RT distributions based on an RT distribution of nouns in BALDEY (Ernestus & Cutler, 2015), we are confident that it is also possible to infer the processing model that best fits human lexical decision data. This validates our methodological approach.

## Appendix C

For determining the parsing penalty prior for our models, we assumed that parsing should take at least 1 ms but no longer than 350 ms. Because all models were fitted with log-transformed predictors, we defined the parsing penalty prior on the log scale as well, that is, as  $\Delta p_{log} = \log(e^{Intercept} + \Delta p) - Intercept$ , whereby the Intercept refers to the mean prior of the intercept, being 6.5173 (677 ms) for DLP (see Table 4) and 5.7090 (302 ms) for BALDEY (see Table 5). This implies that, on the log scale, we expect the mean of the parsing penalty to be between 0.0015 (1 ms) and 0.4168 (350 ms) for DLP and between 0.0033 (1 ms) and 0.7704 (350 ms) for BALDEY.

We tested sequences of means, of which some were slightly outside the range of plausible parsing penalties, because this enables us to investigate the model's estimations when provided with an implausible parsing penalty prior. Specifically, we explored normal distributions for the parsing penalty prior with means ranging from 0.001 to 0.801 in 0.02 increments while the corresponding standard deviations (SD) were defined by formula (5).

$$SD_i = mean_i / 4 + 4 / (mean_i + 4). \quad (5)$$

Note that the computed standard deviation tends to be a fourth of the mean for large values of the mean, with a lower bound of 1, which is attained for low values of the mean just above 0 (e.g., mean of 0.001 has a SD of  $0.001/4 + 4/(0.001 + 4) = 1$ ). This is necessary to ensure that priors smaller than 4 ms are not too informative so that updating the prior distribution based on the actual data would be impossible.

## Appendix D

**Table 12** Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 1 (written plural nouns ending in *-en*). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with *FQ*

Estimate	Root-driven model			Form-driven model			Tipping point model		
	Mean	SD	Rhat	Mean	SD	Rhat	Mean	SD	Rhat
Intercept	6.51000	0.02170	1.00	6.52000	0.02140	1.01	6.53000	0.02060	1.01
Trial	0.00814	0.00087	1.00	0.00820	0.00087	1.00	0.00818	0.00087	1.00
Length	-0.00462	0.00176	1.00	-0.00407	0.00176	1.00	-0.00292	0.00177	1.00
Form FQ	-	-	-	-0.02410	0.00054	1.00	-0.02610	0.00060	1.00
Root FQ	-0.02640	0.00046	1.00	-0.02690	0.00041	1.00	-0.02910	0.00045	1.00
Penalty	0.06340 (0.0636)	0.00175	1.00	-	-	-	0.08300 (0.082)	0.00374	1.00

**Table 13** Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 2 (spoken plural nouns ending in *-en*). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with *FQ*

Estimate	Root-driven model			Form-driven model			Tipping point model		
	Mean	SD	Rhat	Mean	SD	Rhat	Mean	SD	Rhat
Intercept	5.85000	0.12200	1.01	5.88000	0.01310	1.01	5.90000	0.13300	1.01
Trial	0.01340	0.00253	1.00	0.01340	0.00261	1.00	0.01340	0.00256	1.00
Duration	0.18700	0.01730	1.01	0.18300	0.01780	1.01	0.18200	0.01780	1.00
Form FQ	-	-	-	-0.00524	0.00143	1.00	-0.00528	0.00145	1.00
Root FQ	-0.00632	0.00121	1.00	-0.00812	0.00111	1.00	-0.00813	0.00113	1.00
Penalty	0.03690 (0.0371)	0.00834	1.00	-	-	-	0.73400 (0.582)	0.04070	1.00

**Table 14** Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 3 (written second/third person singular present tense verb forms ending in *-t*). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with *FQ*

Estimate	Root-driven model			Form-driven model			Tipping point model		
	Mean	SD	Rhat	Mean	SD	Rhat	Mean	SD	Rhat
Intercept	6.42000	0.02300	1.00	6.43000	0.02200	1.01	6.42000	0.02230	1.01
Trial	0.00709	0.00145	1.00	0.00712	0.00146	1.00	0.00714	0.00147	1.00
Length	0.00613	0.00185	1.00	-0.00610	0.00188	1.00	0.00607	0.00186	1.00
Form FQ	-	-	-	-0.01110	0.00106	1.00	-0.01110	0.00106	1.00
Root FQ	-0.00781	0.00083	1.00	-0.00818	0.00074	1.00	-0.00819	0.00073	1.00
Penalty	0.01250 (0.0124)	0.00415	1.00	-	-	-	0.43900 (0.044)	0.26200	1.00

**Table 15** Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 4 (spoken second/third person singular present tense verb forms ending in *-t*). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with *FQ*

Estimate	Root-driven model			Form-driven model			Tipping point model		
	Mean	SD	Rhat	Mean	SD	Rhat	Mean	SD	Rhat
Intercept	5.60000	0.17200	1.01	5.50000	0.16500	1.00	5.49000	0.17000	1.01
Trial	0.01380	0.00282	1.00	0.01380	0.00274	1.00	0.01380	0.00271	1.00
Duration	0.22900	0.02500	1.01	0.24500	0.02440	1.00	0.24500	0.02490	1.00
Form FQ	-	-	-	-0.01230	0.00186	1.00	-0.01230	0.00200	1.00
Root FQ	-0.00593	0.00124	1.00	-0.00472	0.00107	1.00	-0.00469	0.00114	1.00
Penalty	0.00791 (0.0045)	0.00431	1.00	-	-	-	0.75100 (1.15)	0.41100	1.00

**Table 16** Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 5 (written nouns ending in the derivational suffix -heid). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with *FQ*

Estimate	Root-driven model			Form-driven model			Tipping point model		
	Mean	SD	Rhat	Mean	SD	Rhat	Mean	SD	Rhat
Intercept	6.44000	0.02110	1.01	6.45000	0.02110	1.00	6.45000	0.02070	1.01
Trial	0.00769	0.00152	1.00	0.00762	0.00150	1.00	0.00763	0.00151	1.00
Length	0.00568	0.00185	1.00	0.00509	0.00187	1.00	0.00507	0.00187	1.00
Form <i>FQ</i>	-	-	-	-0.00114	0.00144	1.00	-0.01140	0.00148	1.00
Root <i>FQ</i>	-0.01470	0.00085	1.00	-0.00163	0.00083	1.00	-0.01620	0.00083	1.00
Penalty	0.07870 (0.0797)	0.01470	1.00	-	-	-	0.50000 (0.526)	0.22800	1.00

**Table 17** Estimated coefficients for a root-driven, a form-driven, and a tipping point model fitted to the data of Study 6 (spoken nouns ending in the derivational suffix -heid). In the row for the parsing penalty, values in brackets are the highest density value. Frequency is abbreviated with *FQ*

Estimate	Root-driven model			Form-driven model			Tipping point model		
	Mean	SD	Rhat	Mean	SD	Rhat	Mean	SD	Rhat
Intercept	5.5400	0.16900	1.01	5.29000	0.15100	1.00	5.28000	0.14400	1.00
Trial	0.0123	0.00262	1.00	0.01220	0.00256	1.00	0.01240	0.00251	1.00
Duration	0.2350	0.02480	1.01	0.27600	0.02150	1.00	0.27700	0.02060	1.00
Form <i>FQ</i>	-	-	-	-0.01150	0.00178	1.00	-0.01160	0.00179	1.00
Root <i>FQ</i>	-0.00585	0.00116	1.00	-0.00620	0.00120	1.00	0.00618	0.00115	1.00
Penalty	0.02400 (0.237)	0.04410	1.00	-	-	-	0.72200 (0.913)	0.40700	1.00

**Funding information** This research was funded by the Deutsche Forschungsgemeinschaft (Research Unit FOR2373 ‘Spoken Morphology’, grant PL 151/7-2 ‘Central project’ to Ingo Plag, and grant ER 574/1-1 ‘Dutch morphologically complex words: The role of morphology in speech production and comprehension’ to Mirjam Ernestus, Louis ten Bosch and Ingo Plag). Hanno Müller, Louis ten Bosch, and Mirjam Ernestus were financed by the Centre for Language Studies of the Radboud University Nijmegen.

## Declarations

**Competing interests** Mirjam Ernestus is on the editorial board of *Morphology* but had no involvement in the review process. All three authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly

from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40(1), 41–61.
- Amenta, S., & Crepaldi, D. (2012). Morphological processing as we know it: An analytical review of morphological effects in visual word identification. *Frontiers in Psychology*, 3, 232. <https://doi.org/10.3389/fpsyg.2012.00232>.
- Baayen, R. H., & Neijt, A. (1997). Productivity in context: A case study of a Dutch suffix. *Linguistics*, 35(3), 565–588.
- Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX lexical database [CD-ROM]*. Linguistic data consortium.
- Baayen, R. H., Burani, C., & Schreuder, R. (1997a). Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. *Yearbook of Morphology*, 1996, 13–33.
- Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997b). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1), 94–117.
- Baayen, R. H., McQueen, J. M., Dijkstra, T., & Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. *Trends in Linguistics Studies and Monographs*, 151, 355–390.
- Baayen, R. H., Wurm, L. H., & Aycok, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon*, 2(3), 419–463. <https://doi.org/10.1075/ml.2.3.06baa>.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>.
- Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. London: Oxford University Press.
- Bertram, R., Lalne, M., & Karvinen, M. (1999). The role of morphological structure in reading derived words. *Annals of Dyslexia*, 49(1), 41–66.
- Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 26(2), 489–511.
- Beyersmann, E., Cavalli, E., Casalis, S., & Colé, P. (2016). Embedded stem priming effects in prefixed and suffixed pseudowords. *Scientific Studies of Reading*, 20(3), 220–230.
- Blevins, J. P. (2016). *Word and paradigm morphology*. London: Oxford University Press.
- Booij, G. (1995). Inherent versus contextual inflection and the split morphology hypothesis. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1995* (pp. 1–16). Dordrecht: Springer.
- Boudelaa, S., & Marslen-Wilson, W. D. (2011). Productivity and priming: Morphemic decomposition in Arabic. *Language and Cognitive Processes*, 26(4–6), 624–652.
- Brybaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1), 45–50.
- Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, 28(3), 297–332. [https://doi.org/10.1016/0010-0277\(88\)90017-0](https://doi.org/10.1016/0010-0277(88)90017-0).
- Cho, J., Pires, A., & Brennan, J. R. (2024). How large are root and affix priming effects in visual word recognition? Estimation from original data and a Bayesian meta-analysis. *Language, Cognition and Neuroscience*, 39(10), 1291–1309.
- Chuang, Y. Y., & Baayen, R. H. (2021). Discriminative learning and the lexicon: NDL and LDL. Oxford Research Encyclopedia of Linguistics.
- Colé, P., Beauvillain, C., & Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language*, 28(1), 1–13.
- Dawson, N., Rastle, K., & Ricketts, J. (2018). Morphological effects in visual word recognition: Children, adolescents, and adults. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 44(4), 645–654.

- De Grauwe, S., Lemhöfer, K., & Schriefers, H. (2019). Processing derived verbs: The role of motor-relatedness and type of morphological priming. *Language, Cognition and Neuroscience*, 34(8), 973–990.
- De Jong, N. H. (2002). Morphological families in the mental lexicon (Doctoral dissertation). *Katholieke Universiteit Nijmegen*.
- Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology*, 68(8), 1469–1488.
- Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., et al. (2010). The French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42, 488–496. <https://doi.org/10.3758/BRM.42.2.488>.
- Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., et al. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods*, 50, 1285–1307.
- Gabry, J., & Češnovar, R. (2022). *cmdstanr: R interface to 'CmdStan'*. <https://mc-stan.org/cmdstanr/>.
- Geary, J. A., & Ussishkin, A. (2018). Root-letter priming in Maltese visual word recognition. *The Mental Lexicon*, 13(1), 1–25.
- Gonnerman, L. M., Seidenberg, M. S., & Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology. General*, 136, 323–345.
- Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, 103(3), 518–565.
- Gravel, S., Bigman, J. S., Pardo, S. A., Wong, S., & Dulvy, N. K. (2024). Metabolism, population growth, and the fast-slow life history continuum of marine fishes. *Fish and Fisheries*, 25(2), 349–361. <https://doi.org/10.1111/faf.12811>.
- Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior*, 2(1), 1–11. <https://doi.org/10.1007/s42113-018-0011-7>.
- Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, 39, 1041–1070.
- Hay, J., & Baayen, R. H. (2002). Parsing and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 2001* (pp. 203–235). Berlin: Springer.
- Jackendoff, R., & Audring, J. (2020). Relational morphology: A cousin of construction grammar. *Frontiers in Psychology*, 11, 2241.
- Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study with 14,000 Dutch mono- and disyllabic words. *Frontiers in Psychology*, 1, 174.
- Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British lexicon project: Lexical decision data for 28,730 English words. *Behavior Research Methods*, 44(1), 287–304. <https://doi.org/10.3758/s13428-011-0118-4>.
- Luke, S. G., & Christianson, K. (2011). Stem and whole-word frequency effects in the processing of inflected verbs in and out of a sentence context. *Language and Cognitive Processes*, 26(8), 1173–1192. <https://doi.org/10.1080/01690965.2010.510359>.
- Magnusson, M., Andersen, M. R., Jonasson, J., & Vehtari, A. (2020). Leave-one-out cross-validation for Bayesian model comparison in large data. In S. Chiappa & R. Calandra (Eds.), *AISTATS: Vol. 108. Proceedings of the 23rd international conference on artificial intelligence and statistics* (pp. 341–351). PMLR. <https://proceedings.mlr.press/v108/magnusson20a.html>.
- Meunier, F., & Segui, J. (1999). Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language*, 41(3), 327–344. <https://doi.org/10.1006/jmla.1999.2642>.
- Milin, P., Smolka, E., & Feldman, L. B. (2018). Models of lexical access and morphological processing. In G. Libben, M. Goral, & G. Libben (Eds.), *The Oxford handbook of the mental lexicon*, London: Oxford University Press.
- Miller, R. B. (1968). Response time in man-computer conversational transactions. In *Proceedings of the AFIPS '68 fall joint computer conference. Part I* (Vol. 33, pp. 267–277). <https://doi.org/10.1145/1476589.1476628>. AFIPS (Proceedings of the ACM).
- Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76, 165–178.
- Morton, J. (1970). A functional model of memory. In D. A. Norman (Ed.), *Models of human memory* (pp. 203–254). San Diego: Academic Press.
- Müller, H., ten Bosch, L., & Ernestus, M. (2024). The family size effect in visual and auditory word recognition. *Language, Cognition and Neuroscience*, 39(6), 793–814.
- Pain, M. T. G., & Hibbs, A. (2007). Sprint starts and the minimum auditory reaction time. *Journal of Sports Sciences*, 25(1), 79–86. <https://doi.org/10.1080/02640410600718004>.

- Perdijk, K., Schreuder, R., Baayen, R. H., & Verhoeven, L. (2012). Effects of morphological family size for young readers. *British Journal of Developmental Psychology*, *30*(3), 432–445. <https://doi.org/10.1111/j.2044-835X.2011.02053.x>.
- Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes*, *20*, 261–290.
- R Core Team (2021). *R: A language and environment for statistical computing [computer software]*. R foundation for statistical computing. <https://www.R-project.org/>.
- Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, *15*, 507–537.
- Reifegerste, J., Meyer, A. S., & Zwitserlood, P. (2017). Inflectional complexity and experience affect plural processing in younger and older readers of Dutch and German. *Language, Cognition and Neuroscience*, *32*(4), 471–487. <https://doi.org/10.1080/23273798.2016.1247213>.
- Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, *50*(4), 1568–1580.
- Savinova, E., & Maljutina, S. (2021). Evidence for dual-route morphological processing across the lifespan: Data from Russian noun plurals. *Language, Cognition and Neuroscience*, *36*(6), 730–745. <https://doi.org/10.1080/23273798.2021.1879182>.
- Schreuder, R., & Baayen, R. H. (1995). Modeling morphological processing. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (Vol. 2, pp. 257–294). Lawrence Erlbaum Associates.
- Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*(1), 118–139.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.
- Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2008.10296>.
- Solomyak, O., & Marantz, A. (2009). Lexical access in early stages of visual word processing: A single-trial correlational MEG study of heteronym recognition. *Brain and Language*, *108*(3), 191–196. <https://doi.org/10.1016/j.bandl.2008.09.004>.
- Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042–2057.
- Sorensen, T., & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists [Preprint]. arXiv. <https://arxiv.org/abs/1506.06201>.
- Stan Development Team (2022). Stan modeling language users guide and reference manual (Version 2.31) [Computer software manual]. Stan Development Team. <https://mc-stan.org>.
- Stevens, P., & Plaut, D. C. (2022). From decomposition to distribution: An attractor-network account of the lexical-semantic system. *Psychonomic Bulletin & Review*, *29*(6), 1673–1702. <https://doi.org/10.3758/s13423-022-02086-0>.
- Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition*, *7*(4), 263–272.
- Taft, M. (2023). Localist lexical representation of polymorphemic words: The AUSTRAL model. In D. Crepaldi (Ed.), *Linguistic morphology in the mind and brain* (pp. 152–166). London: Routledge. <https://doi.org/10.4324/9781003159759-11>.
- Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 638–647.
- Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, *51*, 1187–1204.
- Vaknin-Nusbaum, V. (2025). Morphological density and reading comprehension in Hebrew novice readers. *Reading & Writing*, *38*(3), 699–721. <https://doi.org/10.1007/s11145-024-10526-7>.
- Vaknin-Nusbaum, V., & Shimron, J. (2011). Hebrew plural inflection: Linear processing in a Semitic language. *The Mental Lexicon*, *6*(2), 197–244. <https://doi.org/10.1075/ml.6.2.01.vak>.
- Vehtari, A. (2020). Cross-validation FAQ [Webpage]. GitHub repository. <https://github.com/stan-dev/loo/blob/HEAD/vignettes/online-only/faq.Rmd>.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>.

- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718. <https://doi.org/10.1214/20-BA1221>.
- Wilder, R. J., Goodwin Davies, A., & Embick, D. (2019). Differences between morphological and repetition priming in auditory lexical decision: Implications for decompositional models. *Cortex*, *116*, 122–142. <https://doi.org/10.1016/j.cortex.2018.10.007>.
- Zou, Y., Tsang, Y. K., Shum, Y. H., & Tse, C. Y. (2023). Full-form vs. combinatorial processing of Chinese compound words: Evidence from mismatch negativity. *International Journal of Psychophysiology*, *187*, 11–19. <https://doi.org/10.1016/j.ijpsycho.2023.02.004>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.