HANNO MÜLLER, LOUIS TEN BOSCH, MIRJAM ERNESTUS

Competing Accounts of the Auditory Family Size Effect: Spreading Activation vs. Discriminative Learning.

Words with more morphologically related words are recognized more quickly. We investigated the cognitive mechanisms underlying these *family size* effects by comparing how well they are accounted for by different models of spoken word recognition. First, we showed that family size effects can partially be accounted for by DIANA, a model of auditory word recognition that does not consider the morphological structures of words. This raises the question of what drives family size effects if it is not morphologically processing. Second, we found that an enriched DIANA, in which we incorporated spreading of activation between morphologically related words, does not better account for auditory family size effects. Third, we showed that discriminative learning accounts somewhat better for family size effects, except for that part that relies on phonological similarity. Together these results suggest that family size effects are more driven by phonological properties of the words than is commonly assumed.

Keywords: auditory word recognition, morphological family size effects, phonological processing, spreading activation, discriminative learning.

1. *Introduction*

A word's family size (FS) is the number of unique word types that includes the word's root. For example, the word *take* has family members including *takeaway, takes*, *took*, and *uptake*. Research has shown that in both visual (e.g., Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004) and auditory (e.g., Müller, ten Bosch, & Ernestus, 2024a), words with larger morphological families are recognized more quickly (henceforth: the FS effect). This article investigates whether and to what extent the spreading activation mechanism (de Jong, Schreuder, & Baayen, 2003) can account for the FS effect in spoken word recognition. The spreading activation is compared to discriminative learning, which has been shown to partially explain the auditory FS effect (Müller et al., 2025).

This section is structured as follows. We first describe the FS effect in more detail (§1.1) and discuss two possible theoretical explanations of this effect (§1.2). We then detail our research questions (§1.3). Subsequently, we describe the human spoken word recognition model DIANA (ten Bosch, Boves, & Ernestus, 2022) in more detail and how we adapted this model to investigate our research questions (§1.4).

1.1 The family size effect

The FS effect has been documented in various studies across multiple languages (e.g., Bertram, Baayen, & Schreuder, 2000; Mulder, Dijkstra, Schreuder & Baayen, 2014) for both reading and listening. Researchers propose that this effect is primarily driven by family members that are semantically similar to the word to be recognized (e.g., Moscoso del Prado Martín et al., 2004). Müller et al. (2024a) observed that family members also contribute more to the FS effect the more they are phonologically similar to that word.

There are systematic differences in how family members affect visual versus auditory word recognition. In visual word recognition, the FS effect is independent of the morphological structure of the word to be recognized. That is, the effect has been documented for the recognition of prefixed (e.g., Moscoso del Prado Martín et al., 2004), simplex (e.g., Mulder, Schreuder, & Dijkstra, 2013), and suffixed words (e.g., Bertram, et al., 2000). In contrast, in auditory word recognition, the FS effect seems restricted to simplex and suffixed words (Müller et al., 2024a).

This difference between the visual and auditory FS effect can be understood by considering how language users perceive words in the two modalities. Readers can see most words at once, because most words fit in the readers' foveal area (approximately ten characters; Legge, Mansfield, & Chung, 2001) or in the combination of the foveal area with the parafoveal preview (e.g., Rayner, 1998). This explains why the FS effect surfaces irrespectively of whether the word's root, which represents the commonality among morphological family members, appears as the first, second, or third morphological constituent. In contrast, in auditory word recognition, words unfold over time, and the recognition process begins as soon as the audio signal starts (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Reinisch, Jesse, & McQueen, 2010). While the speech signal unfolds, listeners consider all possible words stored in their memories that match the incoming speech signal perceived so far, and eliminate those that no longer match (e.g., Marslen-Wilson & Welsh, 1978). As a consequence, prefixes are processed first, followed by roots, and then suffixes. This explains why the FS effect has not been attested for spoken prefixed words: the recognition process is well on its way before the word's root is perceived.

1.2 Theoretical explanations for the family size effect

One explanation for the FS effect is based on the theory of discriminative learning (Widrow & Hoff, 1960; Rescorla & Wagner, 1972), as incorporated in the Discriminative Lexicon Model (DLM, e.g., Chuang & Baayen, 2021). This theory assumes that the connections between elements of a word's form (e.g., sequences of graphemes or sounds) and its meaning are strengthened when they co-occur and that they are weakened when they do not. For instance, the connections between sequences of letters within *take* and the meaning of *take* are strengthened by family members of *take* such as *takeaway* and *takes*, which contain the sequence *take* and have a meaning related to that of *take* ('to move something from one place to another'). Weakening occurs if a grapheme or sound sequence (e.g., /bɪə/) appears in words with different meanings (e.g., *beer*, *beard*) or if a morpheme has multiple meanings (e.g., *bank*). Discriminative learning is expected to account for FS effects because the more family members a word has, the stronger the associations become between this word form's features and its meaning, resulting in quicker recognition when the word is encountered. Mulder and colleagues (2014) showed for visual word recognition that (an earlier version of) DLM can indeed account for at least part of the FS effect. They never tested, however, whether the DLM can account for the complete FS effect.

For auditory word recognition, we found in a previous study (Müller et al., 2025) that the FS effect cannot be completely explained by the DLM. We tested LDL-AURIS (Shafaei-Bajestan, Moradipour-Tari, Uhrig, & Baayen, 2023), which is an auditory version of the DLM. LDL-AURIS takes as its input a word's audio recordings and it produces as its output a vector representing the meaning of that word. Semantic vectors are numerical representations of words' meanings. Words with more similar meanings tend to have more similar semantic vectors, among other reasons, because semantic vectors reflect how often words co-occur with which other words. We derived from LDL-AURIS the predictor semantic density, which measures the average similarity (in terms of cosine similarity) between an input word's semantic vector and the ten most similar semantic vectors in the mental lexicon. With higher predicted semantic density, a stimulus is more likely to be a word, but also less distinct from other words. In a regression model, this predictor explained part of the same variance of auditory lexical decision response times (RTs) as FS does, suggesting that the FS effect may emerge through discriminative learning.

As an alternative to the DLM, the FS effect may be explained with a spreading activation mechanism. A spreading activation mechanism has been implemented in a few models of visual word recognition, for instance, the Morphological Family Resonance Model (MFRM; de Jong et al., 2003). The MRFM's lexicon consists of lemma representations (cf. Levelt, 1989), which are connected with syntactic and semantic representations (cf. Schreuder & Baayen, 1995). Semantic representations coincide with morpheme representations, so that family members are linked to the same semantic representation (e.g., *take*, *takeaway*, *takes*, and *uptake* are all connected to the semantic representation *take*), among other semantic representations. When a lemma representation is activated, it transmits activation to all associated syntactic and semantic representations, and those in turn transmit activation to their associated lemma representations, the original lemma representation included. As a consequence, the activation of an encountered word is increased more if it is associated with more words and thus has more family members. As soon as a lemma representation reaches a predefined threshold value for activation, the corresponding word is recognized. The process of propagating activation from lemma representations to central representations and back again is referred to as a cycle, and the number of cycles it needs for a lemma representation to reach threshold activation level determines how quickly a stimulus is responded to.

De Jong et al. (2003) implemented the MFRM to generate lexical decision RTs for visual lexical decision stimuli. They found that the stimuli's Family Sizes correlated with the generated RTs approximately as well as with observed RTs from a lexical decision experiment. This suggests that the MFRM can (at least partly) account for the visual FS effect. Because it has never been investigated whether similar results can be obtained for auditory lexical decision, it is an open question whether the MFRM can account for the auditory FS effect.

1.3 The present study: Research questions

The present study addresses the general question of which cognitive processes underly the auditory FS effect, focusing on the explanatory power of a spreading activation mechanism in comparison to discriminative learning. Because there is no model of human auditory word recognition available that contains a spreading activation mechanism, we had to create such a model. For doing so, we incorporated a spreading activation mechanism into the process-oriented model DIANA (ten Bosch et al., 2022; see §1.3). We addressed the following research questions.

1. To what extent does DIANA without spreading activation incorporated account for the FS effect?
2. Does augmenting DIANA with a spreading activation mechanism enables DIANA to better account for the auditory FS effect?
3. How well does DIANA account for the FS effect compared to LDL-AURIS?

1.4 DIANA

DIANA is a recent model of human auditory word recognition (see ten Bosch et al., 2022, for an overview). In what follows, we describe the original DIANA, which does not contain a spreading activation mechanism. In §1.4.2, we describe how we incorporated a spreading activation mechanism into DIANA.

DIANA's input is an audio recording, and its output is an ordered list of word hypotheses. These hypotheses and their activation values can be used to predict lexical decision RTs or latencies and amplitudes of EEG components (Bentum, ten Bosch, van den Bosch & Ernestus, 2019). DIANA comprises three main components. The first component, the activation component, calculates activations or probabilities of both word and pseudoword hypotheses at 10 ms time steps during the unfolding of the audio signal (see Fig. 1, white blocks). For doing so, the activation component computes feature vectors that resemble Spectro-Temporal Receptive Fields, which simulate phone activations in the auditory cortex (Mesgarani, Cheung, Johnson, & Chang, 2014). For each time step (t) between signal onset (t = 0) and signal offset, each word hypothesis' probability (P) is computed as a function of the feature vectors computed from 0 to t:

(1) P(hypothesis | feature vectors[0:t])

This approach mirrors the dynamic nature of human auditory word recognition, which is assumed to start as soon as the speech signal is detected.

The second component, the decision component, evaluates the updated word hypotheses generated by the activation component at every time step to determine the winning (pseudo)word hypothesis, and to conclude whether the audio signal represents a word or a pseudoword. The time required for making the decision between word and pseudoword (lexical decision RT) depends on the dynamic competition between word and pseudoword hypotheses.

Ten Bosch et al. (2022) showed that human lexical decision RTs can be well predicted with the decision component considering the entropies of the (pseudo)word hypotheses at word offset, reflecting the principle that resolving ambiguity takes time. Entropy measures the degree of disorder among the hypotheses. The decision component assigns shorter RTs to stimuli with lower entropies at stimulus offset (intercept is a value depending on the experiment, the factor β is a scaling factor in the Hick-Hyman law, n is the number of hypotheses; Proctor & Schneider, 2018):

(2) $$\text{decision time}_{\text{stimulus entropy}} = intercept + \beta \cdot \sum_{i=1}^{n} -p_i \log(p_i)$$

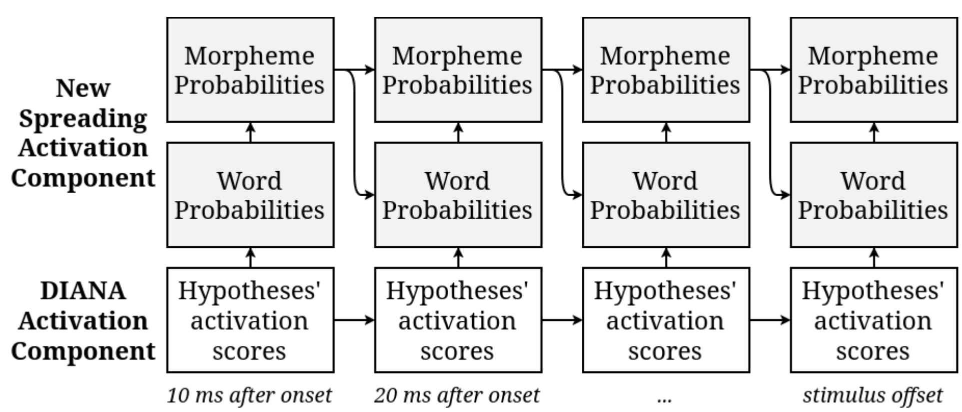
DIANA's third component, the execution component, mimics neural travelling time, reflecting the time it takes for neural activations in the brain to produce a measurable behavior. In lexical decision experiments, this behavior usually is a button press. In DIANA, the execution time is assumed to be constant (i.e., stimulus independent, participant independent).

DIANA's total RT measured from a stimulus' offset depends on the times taken by the decision ('decision time') and the execution ('execution time') as well as on the stimulus duration. Because the longer the stimulus, the more of the processing will have taken place before the end of the stimulus, the factor '*correction offset*' is typically negative:

(3) DIANA RToffset = correction offset · stimulus duration + decision time + execution time

DIANA neither has knowledge about words' morphological structures, nor does it contain spreading activation mechanisms. In addition, setting up DIANA does not involve establishing mechanism that may implicitly provide information about words' morphological structures, such as form-meaning-correspondences in the DLM. We therefore hypothesize that DIANA cannot account for the FS effect. In fact, it may be assumed that words with more family members induce a higher entropy at word offset because the larger a word's FS is, the more probable more family members are, given the audio signal. Consequently, DIANA may be expected to predict longer rather than shorter RTs for larger FS words.

Figure 1 - *Flowchart illustrating how word and morpheme probabilities are updated at each 10ms timestamp in DIANA. White blocks represent DIANA's original activation component. Grey blocks represent the spreading activation mechanism that we incorporated.*



1.4.1 Lexicality score

As mentioned above, the original version of DIANA computes an RT basing the decision time on the stimulus entropy (Hick-Hyman law, see formula 2). This implementation implies that the RT is independent of whether the hypothesis that is selected is mostly followed in probability by words or pseudoword hypotheses. This independence is not entirely plausible for lexical decision tasks, since it can be assumed that lexical decision RTs are (also) based on how likely all real word hypotheses are compared to all pseudoword hypotheses.

We therefore propose an alternative: a decision time based on the *lexicality score*, that is, on a score reflecting the ratio between the probabilities of all real word hypotheses and the probabilities of all pseudoword hypotheses at stimulus offset. The higher the probabilities of the real word hypotheses are in comparison to the probabilities of the pseudoword hypotheses, the shorter the decision time is for a real word. This alternative is again related to entropy, but now applied on the group of words and pseudo words in the candidate hypothesis list, instead of on *individual* hypotheses as was assumed in the 2022-version of DIANA.

The decision time based on lexicality score is the sum of three components (see Eq. 4). The first component is a constant (intercept). The second component is the word / pseudoword ratio multiplied by a negative scaling factor *lex2RT*. This component reflects the facilitation of a "Yes" response for stimuli with high lexicality scores. The third component is the reverse of this ratio multiplied by a negative scaling factor *invlex2RT*. This component reflects the facilitation of a "No" response for stimuli with low lexicality scores.

(4) $\text{decision time}_{\text{lexicality score}} = \text{intercept} + \text{lex2RT} * \text{word ratio} + \text{invlex2RT} * (1 / \text{word ratio})$

In this formula, lex2RT and invlex2RT are the parameters (weight coefficients) that determine to what extent the word ratio and its inverse influence the decision time due to DIANA's decision component.

1.4.2 Augmenting DIANA with a spreading activation mechanism

We implemented into DIANA a spreading activation mechanism that is inspired by the MFRM (de Jong et al., 2003). We incorporated representations for root morphemes in DIANA's lexicon by adding a new morpheme layer, and we set up connections between word representations and the corresponding morpheme representations. The representations of morphologically complex word hypotheses containing more than one morpheme are connected to multiple morpheme representations (e.g., *background* is connected to the morpheme representations *back* and *ground*). Morpheme probabilities are updated every time the connected words' probabilities are updated (i.e., every 10 ms). The morpheme probabilities, in their turn, codetermine the probabilities of morphologically related word hypotheses (see Fig. 1, grey blocks). For more detailed information, see the Appendix.

The spreading activation mechanism as implemented in DIANA involves two meta parameters, *sensitivity* and *decay*, both ranging between 0 and 1. The sensitivity parameter represents the idea that a listener may be more sensitive to the acoustic signal under certain listening conditions (e.g., when hearing with headphones) than under other conditions (e.g., being present at a cocktail party). Mathematically, sensitivity determines the extent to which the probabilities of hypotheses are not only determined by their match with the audio (with a sensitivity value of 1, the probabilities are completely determined by the audio) but also by support from the morphological layer (i.e., top-down expectations. For instance, given a sensitivity of 0.9, if the word *background* has a probability of 0.2 and the morphemes *back* and *ground* have a probability of 0.1 and 0, respectively, the probability of the word *background* will become 0.1 ($0.9 \cdot 0.2 + (1 - 0.9) \cdot (0.1 + 0)$). The decay parameter represents the idea that morphological information slowly disappears when it is not activated anymore. Mathematically, decay determines how much the probability of a morpheme at a time step is determined by its probability at the previous time step and by the probabilities of the connected word hypotheses (with a *decay* value of 1, morpheme probabilities are completely determined by the connected word probabilities of the same time step). For instance, given a decay of 0.5, if the word *running* has a probability of 0.1 and the morpheme *run* has a probability of 0, the probability of the morpheme *run* will become 0.05 ($0.1 \cdot 0.5 + 0 \cdot (1 - 0.5)$).

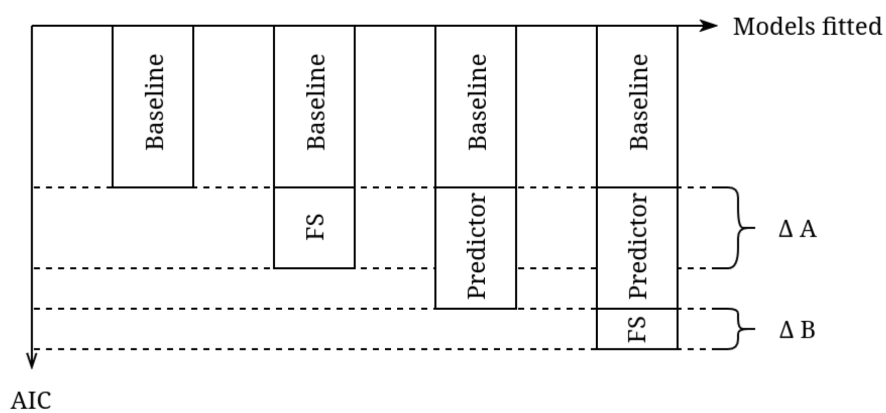
2. *General methods*

We investigated how well different models (the original DIANA with its decision time based on either stimulus entropy or lexicality score, the spreading activation enriched DIANA, and LDL-AURIS) account for the FS effect on the basis of the Biggest Auditory Lexical Decision Experiment Yet (BALDEY; Ernestus & Cutler 2015, see §2.1). To see whether the different models may take into account the contribution of different types of family members, we tested three different family size measures (see §2.2).

We followed the statistical procedure outlined in Fig. 2. We first computed a baseline statistical model predicting the RTs from control variables known to predict auditory lexical decision RTs (see §2.3 and see "Baseline" in Fig. 2). We then enriched the baseline statistical model such that we produced two enriched statistical models for every word recognition model under study: one enriched statistical model also contained a predictor derived from the word recognition model under scrutiny such as lexicality score or semantic density ("Predictor" in Fig. 2), while the other enriched statistical model also contained this same predictor in addition to a family size measure ("FS" in Fig 2; in interaction with morphological structure, because FS interacts with morphological structure; see Müller et al., 2024a). If the latter enriched statistical model better fits the RTs than the former enriched statistical model, the predictor derived from the word recognition model under study does not fully explain the FS effect (see Fig. 2, third and fourth bar).

In that case, we investigated whether the model of word recognition under study accounts for at least part of the FS effect. We compared how much the FS measure (in interaction with morphological structure) improves the model fit when added to the baseline statistical model (see Fig. 2, ΔA) and when added to this baseline model enriched with the predictor from the model under study (see Fig. 2, ΔB). If the FS measure leads to a larger model improvement fit for the baseline model, the model of word recognition at least partially accounts for the FS effect.

Figure 2 - *Schematic illustration of model comparisons that were carried out to investigate whether a model explains portions of the FS effect. This is the case if adding FS to the baseline model yields a greater model improvement in terms of AIC (ΔA) than adding FS to the baseline model plus the predictor from the model of word recognition under study (ΔB).*



The data and the scripts that were used for this study can be downloaded from: https://doi.org/10.34973/71n3-vk61 (link for external review: https://data.ru.nl/login/reviewer-7709611/FWI56KZGPLWZZBW2YXG6TPLCVVTZC45EMVPYSOQ).

2.1 Data

We analyzed the RTs from BALDEY, a Dutch large-scale auditory lexical decision experiment. BALDEY contains RTs from 20 native Dutch speakers on 2,780 spoken Dutch content words and 2,761 pseudowords, both bare roots and inflected and derived words. We only considered accurate responses to real words, excluding compounds, words without morphological analysis in CELEX (Baayen, Piepenbrock, & Gulikers, 1996), and words that do not occur in the Spoken Dutch Corpus (Oostdijk, 2002) and that were therefore not included in training of LDL-AURIS (see §5). Further, we excluded 272 responses (1.71%) that were given before stimulus offset because participants may not have correctly recognized these words, and 24 (0.15%) responses with an encoding error. The resulting dataset contains 15,936 responses in total: 5,908 responses to 322 unique simplex words, 227 responses to 12 unique prefixed words, 8,875 responses to 478 unique suffixed words, and 926 responses to 50 unique words that contain both a prefix and a suffix.

2.2 Three family size measures

We focused on three different FS measures that we have tested in previous research (e.g., Müller et al., 2024a). We computed all these measures on the basis of the CELEX database (Baayen et al., 1996).

Our first FS measure is the word's Classical FS, which is the number of all words with the presented word's root (i.e., all family members). Each family member thus equally contributes to the FS effect.

Our second FS measure is the Semantic FS, which is based on the assumption that a family member contributes the more to the FS effect, the closer its semantic relationship is with the presented word. We quantified the semantic relationship between the presented word and a family member with the help of a distributional semantics model, word2vec (Mikolov, Chen, Corrado, & Dean, 2013), which was trained on more than 600 million messages on Dutch social media, news, blogs, and forums (Nieuwenhuijse, 2018). We computed the cosine of their semantic vectors and normalized the result with min-max scaling so that opposite vectors have a similarity value of 0 and identical vectors have a similarity value of 1. We summed the resulting numbers of all family members. Thus, each family member contributes to the Semantic Family Size according to its semantic similarity with the presented word.

Finally, our third FS measure, Semantic Form Overlap FS, reflects the assumption that family members' contributions to the FS effect are larger with both stronger semantic relationships and larger form overlap with the presented word (Winther Balling & Baayen, 2012). Semantic relationship was calculated on the basis of a distributional semantics model, just as for Semantic FS. Form overlap was based on the words' phonemic representations. That is, the Levenshtein distance between the presented word and a family member was determined, ignoring stress and suffixes;

the outcome was squared (emphasizing differences in the Levenshtein distance) and inverted (so that greater form overlap results in higher membership values), and passed to the exponential function, resulting in a value between 0 and 1 (i.e., the same range as of semantic relationship values). For each family member, this value was multiplied with its semantic similarity with the presented word (as computed for the Semantic FS) and the outcomes for all family members were summed. This FS measure best predicted RTs for lexical decision in Müller et al. (2024a).

Noteworthy, the Classical FS has a perfectly linear relationship with words' numbers of family members. In contrast, the Semantic FS and Semantic Form Overlap FS do not necessarily show linear relationships. For instance, a word with five family members can have a Semantic FS of almost five, when all five family members have strong semantic relationships with this word, but it can also have a Semantic FS of 1, if its family members have weak semantic relationships with this word.

The Classical FS had a mean of 29.69 (SD = 49.83), the Semantic FS a mean of 18.07 (SD = 27.74), and the Semantic Form Overlap FS a mean of 6.65 (SD = 12.96). For the statistical analyses, the measures were log-transformed and subsequently normalized with a z-transformation. The untransformed Classical FS and Semantic FS strongly correlate ($r = .995$). The untransformed Classical FS and Semantic Form Overlap FS correlate less strongly ($r = .782$). The same holds for the untransformed Semantic FS and Semantic Form Overlap FS ($r = .782$).

2.3 Control variables in the baseline model

Our baseline statistical model incorporated four control predictors. The first one, Moving Average Response Time (maRT), reflects a participant's local speed and is the weighted average RT across the 10 previous trials (ten Bosch, Ernestus, & Boves, 2018). The second control predictor, trial number (Trial), accounts for a participant's adaptation over the course of the whole experimental session (e.g., Ernestus & Cutler, 2015). The third control predictor, form frequency (Frequency), captures effects of frequency of occurrence and is derived from CELEX (Baayen et al., 1996). These three control variables were log-transformed and z-transformed to ensure proper scaling and centering.

Our fourth, categorical, control predictor was morphological structure (MorphStr), which categorizes words as "prefixed", "simplex", "suffixed", or "double-affixed" (for word containing both a prefix and a suffix). With this fourth predictor, we can take into account (by means of interactions with the FS measures) that the auditory FS effect differs based on the word's morphological structure (e.g., simplex and suffixed words elicit FS effects whereas prefixed words do not; Müller et al., 2024a).

Table 1 - *Correlation coefficients of the control predictors and FS measures.*

	maRT	Trial	Classial FS	Semantic FS	Semantic Form Overlap FS
Freq	.000	.006	.382	.384	.333
maRT	-	-.003	.005	.005	.005
Trial		-	-.026	-.024	-.019
Classical FS			-	.998	.804
Semantic FS				-	.810

The correlation coefficients between each pair of transformed continuous control variables, as well as between the transformed control variables and either transformed FS measure ranged between .01 and -.03. (see Table 1). The exception was Frequency, which showed a weak correlation with each of the three FS measures (range: 0.33 - 0.38), which is considered unproblematic for the statistical modeling.

2.4 General description of the analysis

All statistical models predicted RTs from word offset. They were implemented as Generalized Additive Mixed Models (GAMMs) in R, version 4.0.5 (R Core Team, 2017), with the *mgcv* package (Wood, 2015). GAMMs were selected over Linear Mixed-Effects Models (e.g., Bates, Mächler, Bolker, & Walker, 2015) due to their ability to accommodate both linear and non-linear effects. According to Baayen, Vasishth, Kliegl & Bates (2017), GAMMs yield a better fit to RTs in BALDEY than Linear Mixed-Effects Models.

Our model fitting followed the approach of Bates et al. (2015). That is, the initial baseline model included a by-participant intercept, by-participant random slopes for all continuous variables, thin plate regression splines for continuous control variables, and a parametric term for the categorical variable morphological structure. In subsequent steps, we simplified the model by removing predictors that were not statistically significant. We checked for concurvity, which occurs when one predictor's effect is largely explained by another predictor. If concurvity was above 0.7, we removed the predictor with the smaller p-value, provided it impacted the significance or shape of the remaining predictor's effect.

Eventually, we compared models using standard methods. We evaluated nested models with the χ^2-test on likelihood scores (e.g., Chuang, Vollmer, Shafaei-Bajestan, Gahl, Hendrix & Baayen, 2021). In case of non-nested models, we considered one model as more likely than another model at the α level of 5% if the former model's

Akaike Information Criterion (AIC; Akaike, 1978) was at least 5.88 points (Wagenmaker and Farrell, 2004) smaller than that of the latter model.

3. Experiment 1

Experiment 1 addressed our first research question, that is, how well DIANA without spreading activation accounts for the FS effect. After presenting the baseline model, we compared which of the measures that can be extracted from DIANA predicts most reliably the lexical decision RTs: decision time based on stimulus entropy (see §1.4) or on lexicality score (see §1.4.1).

3.1 Baseline model

The maximal baseline model that converged is summarized in Table 2. It includes a parametric effect for morphological structure, splines for all other control variables, a by-participant intercept and a by-participant random slope for trial number.

Table 2 - Summary of the baseline model fitted to log(RT from word offset) in BALDEY. For parametric coefficients, the intercept represents morphologically simple words.

A. Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	***p*-value**
(Intercept)	6.263	0.0357	0.036	< .001
MorphStr:Prefixed	-0.091	0.04167	-2.177	0.030
MorphStr:Suffixed	-0.097	0.0104	-9.318	< .001
MorphStr:DoubleAffixed	-0.063	0.022	-2.921	0.004
B. Smooth terms	**Edf**	**Ref.df**	**F**	***p*-value**
s(Frequency)	5.659	6.906	7.704	< .001
s(maRT)	4.798	6.006	289.347	< .001
s(Trial)	5.933	7.105	4.517	< .001
s(Participant)	18.525	19.000	26.586	< .001
s(Trial, Participant)	15.813	19.000	8.911	< .001

The parametric effect for morphological structure is statistically significant for all models tested in the present study. According to the baseline model, simplex words are responded to more slowly than suffixed words ($t = -9.318$, $p < .001$), double-affixed words ($t = -2.921$, $p < .005$), and prefixed words ($t = -2.177$, $p = < .05$). There are no significant differences in RTs between suffixed and prefixed words (t

= 0.143, p = .886), between suffixed and double-affixed words (t = 1.588, p = .112), or between prefixed and double-affixed words (t = 0.607, p = .544).

The other control predictors' effects are as expected, which shows the reliability of our analyses. RTs were longer the lower the participant's local speed, the further the participant was in the experiment (probably as the result of fatigue), and the lower the frequency of the word.

3.2 Stimulus entropy and lexicality score

We first established with which parameter values DIANA's best predicts the lexical decision RTs, both when decision time is based on stimulus entropy and when it is based on lexicality score. We set the execution time to 200 ms (a constant assumed to be independent of stimulus and participant). With a gradient search, we found the following optimal parameters: an intercept of 550 ms, a β of 14, an InvLex2RT of -6.1 and an Lex2RT of -85, and, finally, a correction offset of -0.61.

With these optimal values of DIANA's parameters, we determined whether the model better predicts lexical decision RTs when the decision time is based on stimulus entropy or on lexicality score. Both measures are theoretically based on the probabilities of all word and pseudoword hypotheses. We only took into account the 100 most probable hypotheses at word offset because the other hypotheses hardly affect the decision time, either based on stimulus entropy or lexicality score.

Adding either decision time based on stimulus entropy or lexicality score as a predictor to the baseline model produces a significantly better model ($\chi2(2)$ = 1254.409, p < .001; $\chi2(2)$ = 2547.34, p < .001, respectively). In terms of AIC, the baseline model enriched with decision time based on stimulus entropy (AIC = 26280.6) fits the data worse than the baseline model enriched with decision time based on lexicality score (AIC = 26252.7). The difference in AIC of 26.96 suggests that the latter model is to be preferred.

3.3 How well does DIANA, without spreading activation, account for the family size effect?

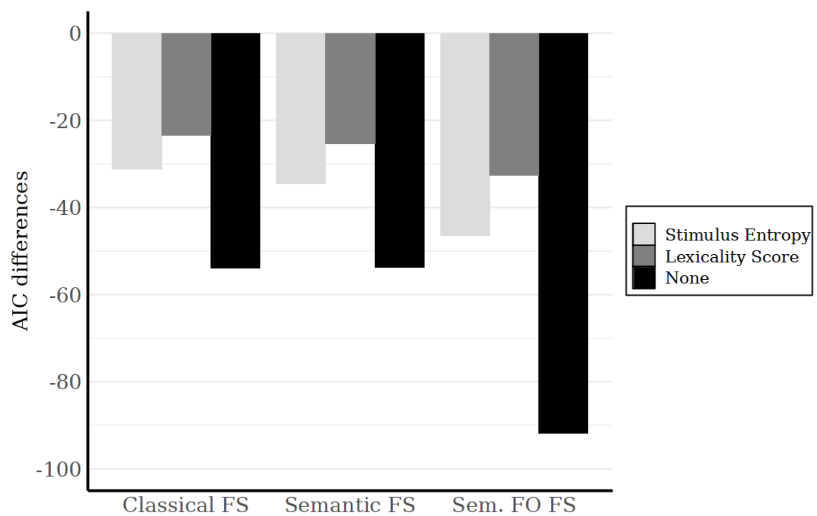
To determine whether DIANA, without spreading activation, completely accounts for FS effects, we determined whether a significantly better statistical model is obtained when any FS measure, in interaction with morphological structure, is added as predictor to the baseline model enriched with DIANA's decision time, based either on stimulus entropy or lexicality score (cf. Fig. 2). Table 3 shows that this is the case. This shows that DIANA does not completely account for the FS effects.

Table 3 - *Comparison of the performance of the baseline model enriched with a DIANA*

measure, either based on stimulus entropy or lexicality score, with the performance of the same model further enriched with any of the three FS measures, in interaction with morphological structure. Sem. FO FS *refers to* Semantic Form Overlap FS.

DIANA measure: decision time based on	**Predictor**	**χ2(8)**	***p*-value**	**AIC Difference**
Stimulus Entropy	Classical FS	15.862	< .001	-31.25
Stimulus Entropy	Semantic FS	17.067	< .001	-34.58
Stimulus Entropy	Sem. FO FS	27.308	< .001	-46.44
Lexicality Score	Classical FS	10.992	< .05	-22.35
Lexicality Score	Semantic FS	11.6597	< .005	-24.26
Lexicality Score	Sem. FO FS	20.413	< .001	-31.57

Figure 3 - *Improvement in AIC points when either FS measure is added to the baseline model and to the baseline model with DIANA's decision time based either on stimulus entropy or Lexicality Score.* Sem. FO FS *refers to* Semantic Form Overlap FS.



In order to see whether DIANA accounts for at least some part of the FS effects, we compared how much the addition of any FS measure improves the baseline model, in terms of AIC, compared to how much it improves the baseline model with DIANA's decision time based on either stimulus entropy or lexicality score as a predictor. The comparison shows that the addition of an FS measure improves the former model more than the latter model (see Fig. 3). This shows that DIANA accounts for part of the FS effects, even without spreading activation incorporated.

The last column of Table 3 suggests that adding either FS measure results in a smaller model improvement, in terms of AIC when, in the statistical model, DIANA is represented by decision time based on lexicality score than on stimulus entropy. This shows that the measure based on lexicality score better accounts for FS effects than the one based on stimulus entropy. For this reason, Experiments 2 and 3 will only consider decision times based on lexicality score.

4. *Experiment 2*

Experiment 2 investigated our second research question, that is, whether a spreading activation mechanism enables DIANA to more accurately predict RTs and to better account for the FS effects.

4.1 Methods

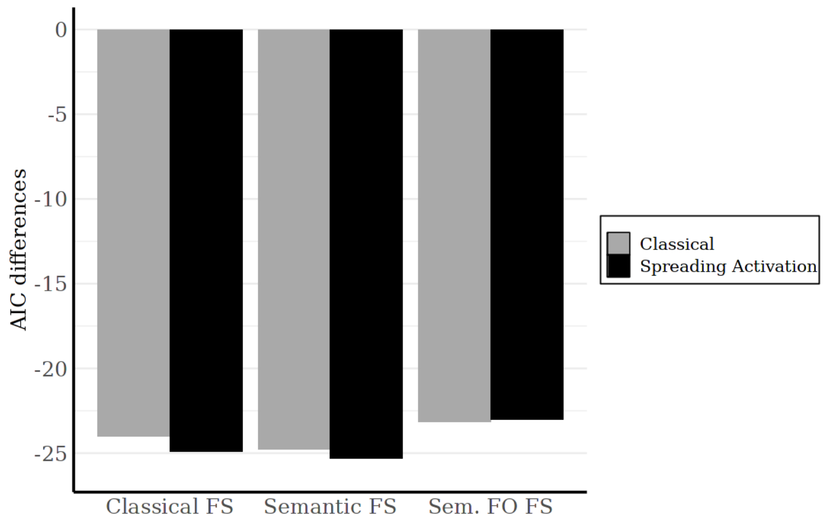
We enriched DIANA with spreading activation as described in §1.4.2 and in the Appendix. We derived the root morphemes and the morpheme-word relationships from CELEX (Baayen et al., 1996), which is the largest database for Dutch words with morphological annotations known to us. We obtained the optimal values for the parameters sensitivity and decay by determining with a grid search with which values DIANA with spreading activation mechanism best predicts the FS effects in our data.

4.2 Results

The optimal values of sensitivity and decay depend on whether the spreading activation mechanism is optimized for accounting for Classical FS and Semantic FS (sensitivity = 0.05, decay = 0.825) versus Semantic Form Overlap FS (sensitivity = 0.75, decay = 0.9). While these optimal values hardly differ for decay, they greatly do for sensitivity (0.05 versus 0.75).

The statistical baseline model with DIANA's decision time (based on lexicality score) as additional predictor is equally improved with the addition of an FS measure, in interaction with morphological structure, when DIANA does versus does not incorporate spreading activation (see Fig. 4). Numerically, incorporating spreading activation in DIANA even leads to a larger contribution of Classical FS (ΔAIC = 0.9) and Semantic FS (ΔAIC = 0.5) to explaining the variance in the data, and hardly any difference in the contribution of Semantic Form Overlap FS (ΔAIC = -0.1). As discussed in Section 2.4, these differences are statistically insignificant.

Figure 4 - *AIC improvement (y-axis) yielded by adding one of the three FS measures to either the baseline model enriched with lexicality score from the Classical DIANA or DIANA with spreading activation.* Sem. FO FS *refers to* Semantic Form Overlap FS.



5. Experiment 3

Experiment 3 investigated our third research question, that is, how well DIANA accounts for the FS effect relative to LDL-AURIS. Experiment 1 already showed how much the predictor lexicality score derived from DIANA accounts for the FS effects. For determining how much LDL-AURIS accounts for the FS effects in the same data set, we first had to train an LDL-AURIS model, so that we could derive LDL-AURIS prediction for each word in the dataset, semantic density, from this model.

We trained LDL-AURIS as we did in Müller et al., (2024). We used julia (Bezanson, Edelman, Karpinski & Shah, 2017) and the package JudiLing (Luo, Chuang, & Baayen,. 2020), setting all parameters exactly as Shafaei-Bajestan and colleagues (2023) did. The input training data consisted of the word tokens from Component O (read-aloud speech recordings from Dutch native speakers) of the Spoken Dutch Corpus (Oostdijk, 2000). Because read-aloud speech is typically clearly pronounced, the recordings in Component O are similar to the stimuli of BALDEY. We sliced out word tokens from their acoustic context, drawing on the segmentations of the Spoken Dutch Corpus. We removed mispronounced, incomplete, and unintelligible word tokens, resulting in a dataset of 550,688 word tokens (39,278 word types). As output training data, we provided LDL-AURIS with each word token's semantic vector. Semantic vectors were taken from the Dutch distributional semantics model that we also used to determine Semantics FS and Semantic Form Overlap FS (see §2.2).

5.1 Results

Adding the predictor semantic density from LDL-AURIS, in interaction with morphological structure, to the baseline model produces a significantly better model ($\chi2(8) = 1253.630$, $p < .001$). However, in terms of AIC, the baseline model enriched with semantic density (AIC = 26297.3) fits the data worse than the baseline model enriched with lexicality score (AIC = 26252.7, see Experiment 1). The difference in AIC of 44.6 suggests that DIANA predicts the RTs better than LDL-AURIS does.

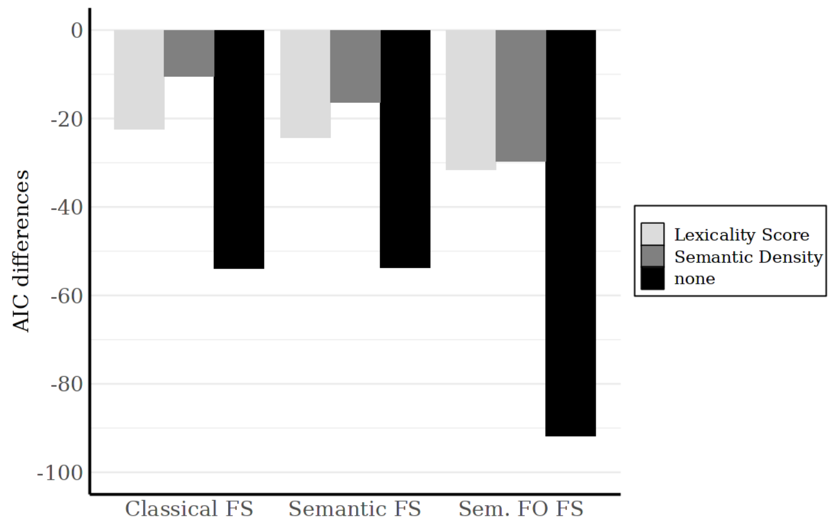
The baseline model with semantic density as additional predictor is improved by the addition of any FS measure, in interaction with morphological structure (see Table 4). This result is in line with our results from Müller et al., (2024), showing that, like DIANA, LDL-AURIS does not completely account for FS effects.

Table 4. Comparison of, on the one hand, the baseline model enriched with semantic density and, on the other hand, the same model further enriched with either of the three FS measures (in interaction with morphological structure). Sem. FO FS *refers to* Semantic Form Overlap FS

Predictor	χ2(8)	*p*-value	AIC Difference
FS	23.958	< .05	-10.552
Semantic FS	27.398	< .05	-16.414
Semantic FO FS	50.260	< .001	-29.606

The improvement in AIC resulting from either FS measure is smaller when this predictor is added – in interaction with morphological structure – to a statistical model with semantic density as a predictor than to a statistical model without this predictor, as can be seen in Fig. 5. This result is also in line with our result in Müller et al., (2024), showing that LDL-AURIS accounts for parts of the FS effects.

Figure 5 - *Improvement in AIC points (y-axis) when adding either FS measure to the baseline model or the baseline model with lexicality score (see Experiment 1) or semantic density (in interaction with morphological structure).* Sem. FO FS *refers to* Semantic Form Overlap FS.



Importantly, the differences in AIC provided in Tables 3 and 4 (last columns) suggest that adding any FS measure to the baseline model results in a smaller model improvement in terms of AIC when the baseline model already contains semantic density (instead of lexicality score). This difference is significant for Classical FS (ΔAIC = 11.796) and Semantic FS (ΔAIC = 7.844), but not for Semantic Form Overlap FS (ΔAIC = 1.961). LDL-AURIS thus seems to better explain Classical FS and Semantic FS than DIANA (without spreading activation).

6. Discussion

This study addressed the question whether the auditory family size (FS) effect can be accounted for by the mechanism of spreading activation, which spreads activation from activated words to their morphologically related words, via shared morphemes, and back (de Jong et al., 2003). We tested the spreading activation mechanism by incorporating it in the spoken human word recognition model DIANA (ten Bosch et al., 2022). We first investigated to what extent DIANA without spreading activation accounts for FS effects. We compared the performance of DIANA's original measure for decision time based on stimulus entropy with the performance of a new one based on lexicality score. Subsequently, we investigated whether

DIANA augmented with a spreading activation mechanism better accounts for FS effects. Third, we investigated how well DIANA accounts for the FS effect in comparison to the Discriminative Lexicon Model (DLM; e.g., Chuang & Baayen, 2021). We based our research on the auditory lexical decision reaction times (RTs) from BALDEY (Ernestus & Cutler, 2015).

6.1 DIANA partly accounts for the family size effect

Our results indicate that DIANA without spreading activation accounts for parts of the FS effects. This result is unexpected because this original version of DIANA does not incorporate any morphological knowledge. This finding suggests that the FS effect is not solely driven by the processing of morphological information.

One possible explanation for this unexpected finding is that morphologically related words are typically phonologically similar, and phonologically similar words may affect each other's recognition. However, a higher number of phonologically related words (or continuation forms) in the lexicon typically leads to more uncertainty in the system about which word was exactly uttered and therefore (according to Hick's law) to longer RTs. In contrast, a higher number of morphologically related words leads to shorter RTs.

Another possible explanation is that family size is correlated with word frequency (e.g., Baayen, Tweedie, & Schreuder, 2002), which would be a proxy for syntactic and morphological co-occurrence probabilities (McDonald & Shillcock, 2001) and these syntactic and morphological co-occurrence probabilities are incorporated into DIANA, by means of a language model. Note that, although DIANA can incorporate frequency information (with frequency weightings of lexical hypotheses; ten Bosch, et al., 2022), we refrained from this possibility. We took frequency effects into account by including frequency as a control variable in the regression models.

In sum, this study suggests that FS effects are not just the result of morphological processing. Future research is needed that reveals exactly why a model of speech processing that does not involve morphological processing nevertheless accounts for part of the FS effects.

We tested how well DIANA accounts for FS effects by studying two measures that can be derived from DIANA: decision time based on stimulus entropy, which was used in previous studies (see ten Bosch et al., 2022) and decision time based on lexicality score, which we defined in the present study and which may be argued to better reflect the processes underlying lexical decision. While both measures predict the auditory lexical decision RTs analyzed in the present study to some extent and while both partly explain the FS effect, the decision time based on lexicality score performs better in both respects. We therefore addressed the remaining research questions for DIANA's decision time based on lexicality score.

6.2 Insights into theoretical accounts of the auditory FS effect

As mentioned above, we tested whether spreading activation between words and morphemes can account for the complete FS effects. We implemented in DIANA morpheme representations and a mechanism that enables word representations to activate morpheme representations and vice versa. This version of DIANA did not account for a larger part of the FS effects than the original version of DIANA. This suggests that a spreading mechanism does not help to explain the FS effect.

Our previous work suggests that discriminative learning can at least partly account for auditory FS effects (Müller et al., 2024). We investigated whether a discriminative learning, as implemented in LDL-AURIS (Shafaei-Bajestan et al., 2023) better accounts for the auditory FS effect than DIANA does, by testing the two models on the basis of the same data set. Interestingly, we found that the LDL-AURIS yields a worse fit to our RTs than DIANA, but better accounts for the Classical FS effect and Semantic FS effect than DIANA. One explanation for this finding is that LDL-AURIS is based on a theory about how morphological information is processed, which is informed by a great body of research (e.g., Baayen, Milin, Đurđević, Hendrix & Marelli, 2011; Baayen & Smolka, 2020; Tomaschek, Plag, Ernestus & Baayen, 2021), while the development of DIANA is based on abundant research on speech processing.

DIANA and LDL-AURIS account for the effect of the Semantic Form Overlap FS to the same extent. One explanation for why DIANA does not perform worse – as it does for the effects of the Classical FS and the Semantic FS – may be that Semantic Form Overlap FS explicitly takes into account how well each word stored in the lexicon matches the presented audio signal, which forms the core activity of DIANA. Given that DIANA does not incorporate morphological processing, and given that Semantic Form Overlap FS best accounts for lexical decision RTs (Müller et al., 2024a), this result suggests that family size effects may be more driven by phonological properties of the words than is commonly assumed.

6.3 The role of alternative model implementations

The results obtained in the present study may raise the question to what extent they depend on how exactly we formulated the spreading activation mechanism in DIANA. As mentioned in §1.4.2, for theoretical reasons, our implementation of the spreading activation mechanism deviated in two aspects from the implementation of the spreading activation mechanism in the MFRM by de Jong and colleagues (2003). First, while de Jong and colleagues implemented a spreading activation mechanism between lemma representations on the one hand and representations for semantics, syntactical classes, and affix representations on the other hand, for theoretical reasons, we implemented in DIANA a spreading activation mechanism between word and morpheme representations. Second, we did not incorporate cycles to compute reaction times. Alternative implementations of the spreading activation mechanisms may better account for auditory FS effects.

When implementing the spreading activation mechanism, we realized that it is unclear how exactly activation should be propagated from words to morphemes and vice versa. In our implementation (see the Appendix), we assumed that a word transmits its probability equally to all its morphemes. For instance, if the word *butterfly* has a probability of 0.5, the morphemes *butter* and *fly* each have a probability of 0.5 too (that is, if their probabilities are not co-determined by their probabilities at the previous time step). In contrast, we assumed that how strongly a morpheme's probability affects a word's probability depends on the word's number of morphemes: the probability propagated by a morpheme is its probability divided by the number of morphemes in the word. For instance, if the morpheme *butter* has a probability of 0.6, the word *butter* receives a probability of 0.6 from the spreading activation mechanism, whereas the word *butterfly*, which has two morphemes, only receives half as much support. There are obviously alternative assumptions possible, and unfortunately, it is difficult to favor one assumption above another based on existing theories. Future research could investigate whether a spreading activation mechanism based on (slightly) different assumptions better accounts for FS effects.

Finally, our implementation of the spreading activation mechanism takes into account the relevance of form overlap between the presented word and the family member for the FS effect (e.g., Winther-Balling & Baayen, 2012), by means of the sensitivity parameter: the higher this parameter, the less family members that do not substantially match the audio signal contribute to spreading of activation. In contrast, this implementation does not take into account the relevance of semantic similarity between the presented word and the family member for the FS effect (e.g., Moscoso del Prado Martín et al., 2004). Future research could try to bring semantic similarity into the spreading activation mechanism.

7. Conclusion

The present study shows that auditory FS effects are partly explained by the original version of DIANA, which does not include morphological processing. This raises questions about the exact nature of FS effects. Augmenting DIANA with connections between word and morpheme representations does not enable DIANA to explain a greater portion of the family size effects and decreases how accurately it accounts for lexical decision RTs. A spreading activation mechanism, at least as we implemented it, thus cannot account for auditory FS effects. The principle of discriminative learning as implemented in LDL-AURIS accounts for larger parts of the FS effects. However, when the family size effects are based both on semantic and form similarity, the two models perform equally well. Together this suggests that explaining the auditory family size effect may require the combinations of two theories: a theory of how morphological information is processed and a theory of audio signal processing.

Acknowledgements

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).

References

ALLOPENNA, P. D., MAGNUSON, J. S. & TANENHAUS, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, 38(4), 419-439.

BAAYEN, R. H., MILIN, P., ĐURĐEVIĆ, D. F., HENDRIX, P. & MARELLI, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*(3), 438–481.

BAAYEN, R. H., PIEPENBROCK, R. & GULIKERS, L. (1996). The celex lexical database (cd-rom). *Linguistic Data Consortium*.

BAAYEN, R. H. & SMOLKA, E. (2020). Modeling morphological priming in German with naive discriminative learning. *Frontiers in Communication, 5*.

BAAYEN, R. H., VASISHTH, S., KLIEGL, R. & BATES, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234.

BAAYEN, R. H., TWEEDIE, F. J. & SCHREUDER, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language*, *81*(1-3), 55-65.

BATES, D., MÄCHLER, M., BOLKER, B. & WALKER, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01.

BENTUM, M., TEN BOSCH, L., VAN DEN BOSCH, A. & ERNESTUS, M. (2019, September). Listening with great expectations: An investigation of word form anticipations in naturalistic speech. In *Interspeech 2019: 20th Annual Conference of the International Speech Communication Association* (pp. 2265-2269).

BERTRAM, R., BAAYEN, R. H. & SCHREUDER, R. (2000). Effects of family size for complex words. *Journal of Memory and Language, 42*(3), 390–405.

BEZANSON, J., EDELMAN, A., KARPINSKI S. & SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM review*, *59*(1), 65-98.

CHUANG, Y. Y. & BAAYEN, R. H. (2021). *Discriminative learning and the lexicon: NDL and LDL. Oxford research encyclopedia of linguistics*. DOI: https://doi.org/10.1093/acrefore/9780199384655.013.375

CHUANG, Y. Y., VOLLMER, M. L., SHAFAEI-BAJESTAN, E., GAHL, S., HENDRIX, P. & BAAYEN, R. H. (2021). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods, 53*, 945-976.

DE JONG, N. H., SCHREUDER, R. & BAAYEN, R. H. (2003). Morphological resonance in the mental lexicon. In R. H. Baayen & R. Schreuder (Eds.), *Morphological Structure in Language Processing* (pp. 65–88). Berlin, Germany: Mouton de Gruyter.

ERNESTUS, M. & CUTLER, A. (2015). Baldey: A database of auditory lexical decisions. *Quarterly Journal of Experimental Psychology, 68*, 1469–1488.

LEGGE, G. E., MANSFIELD, J. S. & CHUNG, S. T. (2001). Psychophysics of reading: XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision research, 41*(6), 725-743.

LEVELT, W. J. (1989). Speaking: From intention to articulation. MIT press.

LUO, X., CHUANG, Y. Y. & BAAYEN, R. H. (2020). JudiLing: an implementation in Julia of Linear Discriminative Learning algorithms for language model. Eberhard Karls Universität Tübingen, Seminar für Sprachwissenschaft.

MARSLEN-WILSON, W. D. & WELSH, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*(1), 29-63.

MCDONALD, S. A. & SHILLCOCK, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295-322.

MESGARANI, N., CHEUNG, C., JOHNSON, K. & CHANG, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006-1010.

MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

MOSCOSO DEL PRADO MARTÍN, F., BERTRAM, R., HÄIKIÖ, T., SCHREUDER, R. & BAAYEN, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1271–1278.

MULDER, K., DIJKSTRA, T., SCHREUDER, R. & BAAYEN, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language, 72*, 59–84.

MULDER, K., SCHREUDER, R. & DIJKSTRA, T. (2013). Morphological family size effects in L1 and L2 processing: An electrophysiological study. *Language and Cognitive Processes, 28*(7), 1004–1035.

MÜLLER, H., TEN BOSCH, L. & ERNESTUS, M. (2024). Can the Discriminative Lexicon Model account for the family size effect in auditory word recognition? *Nota Bene, 1*(2), 176-192.

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.

NENADIĆ, F., TUCKER, B. & TEN BOSCH, L. (2023). Computational modeling of an auditory lexical decision experiment using DIANA. *Language and Speech, 66*(3), 564–605.

NIEUWENHUIJSE, A. (2018). Dutch word2vec model. Retrieved 2022-02-10, from https://github.com/coosto/dutch-word-embeddings

OOSTDIJK, N. (2002). The design of the Spoken Dutch Corpus. In P. Peters, P. Collins, & A. Smith, *New Frontiers of Corpus Research* (pp. 105–112). Amsterdam: Rodopi.

PROCTOR, R. W. & SCHNEIDER, D. W. (2018). Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, *71*(6), 1281-1299.

R CORE TEAM. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

RAYNER, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422.

REINISCH, E., JESSE, A., & MCQUEEN, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology, 63*(4), 772-783.

RESCORLA, R. A. & WAGNER, A. R. (1972). In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts, chapter A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement.

SCHREUDER, R. & BAAYEN, R. H. (1995). Modeling Morphological Processing. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing* (pp. 257–294). New York, USA: Psychology Press.

SHAFAEI-BAJESTAN, E., MORADIPOUR-TARI, M., UHRIG, P. & BAAYEN, R. H. (2023). LDL-AURIS: A computational model, grounded in error-driven learning, for the comprehension of single spoken words. *Language, Cognition and Neuroscience*, *38*(4), 509–536.

TEN BOSCH, L., BOVES, L. & ERNESTUS, M. (2022). DIANA, a process-oriented model of human auditory word recognition. *Brain Sciences, 12*(5), 681.

TEN BOSCH, L., BOVES, L., TUCKER, B. & ERNESTUS, M. (2015). DIANA: towards computational modeling reaction times in lexical decision in North American English. *Proceedings of Interspeech 2015*, 1576–1580.

TEN BOSCH, L., ERNESTUS, M.& BOVES, L. (2018). Analyzing Reaction Time Sequences from Human Participants in Auditory Experiments. *Proceedings of Interspeech 2018: The 19th Annual Conference of the International Speech Communication Association*, pages 971-975.

TOMASCHEK, F., PLAG, I., ERNESTUS, M. & BAAYEN, R. H. (2021). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics, 57*(1), 123-161.

WAGENMAKERS, E. J., FARRELL, S. & RATCLIFF, R. (2004). Estimation and interpretation of 1/fα noise in human cognition. *Psychonomic bulletin & review*, 11(4), 579-615.

WIDROW, B. & HOFF, M. (1960). Adaptive switching circuits. In *1960 WESCON convention record part IV*.

WINTHER BALLING, L. & BAAYEN, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. Cognition, 125(1), 80-106.

WOOD, S. (2015). Package 'mgcv'. R package version, 1, 29.

Appendix: Spreading activation in DIANA

As explained in §1.4.2 and as illustrated in Fig.1, we augmented DIANA with a) morpheme representations in the lexicon, b) information on the relationships between morpheme and word representations, and c) a mechanism that updates morpheme probabilities based on word probabilities and, vice versa, that updates word probabilities based on morpheme probabilities, taking word-morpheme-relationships into account. Here, we illustrate how we did so, on the basis of a case in which a 20 ms long word token of *three* is uttered, and the lexicon just contains three words (*three*, *threefold*, *fold*) and two morphemes (*three*, *fold*). Moreover, for simplicity, we assume that in the course of the word recognition process, only three pseudowords are activated (*thu*, *tri* and *sri*).

The word, morpheme, and pseudoword hypotheses form matrices (W, M, and N, respectively, see A.1):

(A.1) $$\mathbf{W} = \begin{array}{l} \begin{pmatrix} W_{1,1} \\ W_{2,1} \\ W_{3,1} \end{pmatrix} \begin{array}{l} three \\ threefold \\ fold \end{array} \end{array}, \quad \mathbf{M} = \begin{pmatrix} M_{1,1} \\ M_{2,1} \end{pmatrix} \begin{array}{l} three \\ fold \end{array}, \quad \mathbf{N} = \begin{pmatrix} N_{1,1} \\ N_{2,1} \\ N_{3,1} \end{pmatrix} \begin{array}{l} thu \\ tri \\ sri \end{array}$$

The morphological relationships between the word hypotheses and the morphemes are represented by matrix L, with w columns, corresponding to the words, and m rows, corresponding to the morphemes. Whenever a morpheme is part of a given word, the cell identified by this word and this morpheme has a value of 1, and otherwise 0 (see A.2).

(A.2)

$$\mathbf{L} = \begin{array}{c} \\ \\ \end{array}\begin{array}{c} \begin{array}{ccc} \textit{three} & \textit{threefold} & \textit{fold} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \textit{three} \\ \textit{fold} \end{array}$$

Column normalization ensures that the values in one column sum up to one (see A.3). This is necessary so that, through spreading activation, words consisting of multiple morphemes receive the same maximum amount of activation as words consisting of fewer morphemes.

(A.3)

$$\mathbf{L} \odot \left(\sum_{i=1}^{m} L_{i,w}\right)^{-1} = \begin{array}{c} \begin{array}{ccc} \textit{three} & \textit{threefold} & \textit{fold} \end{array} \\ \begin{pmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \textit{three} \\ \textit{fold} \end{array}$$

The word (WDIANA) and pseudoword (NDIANA) probabilities at t_{10} and t_{20} as provided by DIANA's activation component just on the basis of the audio signal are shown in Equations (A.4) and (A.5), respectively. See ten Bosch et al., (2022) for how these probabilities are computed.

(A.4)

$$\mathbf{WDIANA^{10}} = \begin{pmatrix} 0.2 \\ 0.1 \\ 0 \end{pmatrix} \begin{array}{l} \textit{three} \\ \textit{threefold} \\ \textit{fold} \end{array}, \quad \mathbf{NDIANA^{10}} = \begin{pmatrix} 0.3 \\ 0.2 \\ 0.2 \end{pmatrix} \begin{array}{l} \textit{thu} \\ \textit{tri} \\ \textit{sri} \end{array}$$

(A.5)

$$\mathbf{WDIANA^{20}} = \begin{pmatrix} 0.4 \\ 0.2 \\ 0 \end{pmatrix} \begin{array}{l} \textit{three} \\ \textit{threefold} \\ \textit{fold} \end{array}, \quad \mathbf{NDIANA^{20}} = \begin{pmatrix} 0.2 \\ 0.1 \\ 0.1 \end{pmatrix} \begin{array}{l} \textit{thu} \\ \textit{tri} \\ \textit{sri} \end{array}$$

Every time these word and pseudoword probabilities are updated on the basis of the acoustic signal, spreading activation takes place. That is, the bottom-up word probabilities given the acoustic signal $DIANA^{t}$ and the top-down word probabilities from spreading activation M^{t-1} are combined into the matrix W^{t}, by weighing the contribution of both types of probabilities as a function of the parameter *sensitivity s*, which lies in the interval between 0 and 1, (see A.6).

(A.6)

$$\mathbf{W^{t}} = s \times \mathbf{WDIANA^{t}} + (1 - s) \times \mathbf{M^{t-1}} \times \mathbf{L} \odot \left(\sum_{i=1}^{m} L_{i,w}\right)^{-1}$$

To ensure that the summed pseudoword and word probabilities amount to 1, both probabilities are normalized. After the word probabilities have been updated, the morpheme probabilities at the current time stamp t (M^{t}) are also updated on the basis of 1) the morpheme probabilities from the previous time step M^{t-1}, 2) the word probabilities at the current time step W^{t}, and 3) the *decay* parameter *d*, which has a value between 0 and 1, (see A.7).

(A.7)
$$\mathbf{M^t} = d \times \mathbf{M^{t-1}} + (1 - d) \times \mathbf{W^t} \times \mathbf{L}^{\top}$$

Hanno Müller – Centre for Language Studies/Faculty of Arts, Radboud University, The Netherlands

https://orcid.org/0009-0007-1784-5148

Mueller, Hanno <Hanno.Mueller@hpi.de>

Louis ten Bosch – Centre for Language Studies/Faculty of Arts, Radboud University, The Netherlands

https://orcid.org/0000-0002-0152-9024

Louis.tenbosch@ru.nl

Mirjam Ernestus – Centre for Language Studies/Faculty of Arts, Radboud University, The
https://orcid.org/0000-0002-1853-0750

Mirjam.ernestus@ru.nl