INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The recognition of compounds: a computational account

L. ten Bosch[1], *L. Boves*[1], *M. Ernestus*[1,2]
[1]Radboud University Nijmegen, NL
[2]Max Planck Institute for Psycholinguistics
{l.tenbosch, l.boves, m.ernestus}@let.ru.nl

Abstract

This paper investigates the processes in comprehending spoken noun-noun compounds, using data from the BALDEY database. BALDEY contains lexicality judgments and reaction times (RTs) for Dutch stimuli for which also linguistic information is included. Two different approaches are combined. The first is based on regression by Dynamic Survival Analysis, which models decisions and RTs as a consequence of the fact that a cumulative density function exceeds some threshold. The parameters of that function are estimated from the observed RT data. The second approach is based on DIANA, a process-oriented computational model of human word comprehension, which simulates the comprehension process with the acoustic stimulus as input. DIANA gives the identity and the number of the word candidates that are activated at each 10 ms time step.

Both approaches show how the processes involved in comprehending compounds change during a stimulus. Survival Analysis shows that the impact of word duration varies during the course of a stimulus. The density of word and non-word hypotheses in DIANA shows a corresponding pattern with different regimes. We show how the approaches complement each other, and discuss additional ways in which data and process models can be combined.

Index Terms: Dynamic Survival, human word comprehension, computational model, compounds

1. Introduction

A large body of psycholinguistic behavioral data, and MEG data [1], in combination with findings from computational models such as Shortlist-B [2], have led to substantial insight in how humans are able to identify words in the acoustic speech stream in which cues for word boundaries are not or very weakly present [3, 4]. In conversational speech, the search for the most plausible word sequence is at least partly supported by the expectations based on contextual and pragmatic information, but these top-down cues are weak or absent in the case of isolated words. This raises the question how participants in an auditory lexical decision experiment deal with word segmentation in long stimuli, and to what extent behavioral measures such as reaction times can uncover underlying cognitive processes.

In recent years, many studies have dealt with the word segmentation problem. An often-cited example in English is 'ship inquiry' [5]. In British English, this utterance has initial overlap with the pronunciation of many other words such as ship, shipping, choir, ink, inquire, inquiry, why, wire, wiry. Activated embedded words slow down the reaction time in a lexical decision task [3]. The competition between embedded candidate words can become quite complex, especially if the remainder has initial overlap with a real word (e.g., [6, 4]).

Several word competition effects can be explained by existing computational models of human word comprehension. Most models (e.g., TRACE [7], the Cohort model [3], and Shortlist-B [2]) assume that words that are compatible with the input enter a competition which each other; in Shortlist-B this competition takes place between sequences of candidate words. All these computational models assume, in some way, that the input speech can be represented in terms of a sequence of phone-like symbols or (as in TRACE) as a sequence of feature vectors that may act as place holders for phones. In these models, the output is defined by the match between signal and word representations on a symbolic level. While a symbolic account is usually insightful, it sidesteps part of the problem a listener is actually confronted with: how to convert the continuous signal into a sequence of discrete units. SpeM ([8]) addresses this by assuming a prelexical component which maps the speech signal into a phone lattice that is then used as input for a lattice-based word search, but it cannot deal properly with mapping errors.

In this paper we address decoding of long spoken words with two complementary approaches, both with the acoustic signal as input. Both experiments use the same set of word stimuli. The first approach is based on statistical modeling of reaction times (RTs). This analysis is based on reaction time data on compounds in BALDEY [9], by using competing risks survival analysis by Generalized Additive Models [10, 11]. In risk analysis, the regression is not explaining the RTs themselves; instead, the parameters of the underlying decision process are modeled from which the RTs are a consequence. In combination with regression splines, this allows to compute the regression coefficients changing as function of, e.g., time from stimulus onset. This makes it possible to discover systematic changes between early and late effects.

The second approach is based on DIANA. DIANA [12, 13, 14, 15] is a recently developed computational model of human speech comprehension that takes the acoustic waveform as input. For each 10 ms time step from stimulus onset, DIANA produces a sorted list of activated word sequences and their activation scores. DIANA does not assume a symbolic prelexical representation to facilitate the match between acoustic input and lexical representation, but directly maps the signal onto its word representations [14].

Fig. 1 provides an example of the number of word embeddings in Dutch, by using the phonetic transcriptions of words. It displays the average ($\pm\sigma$) number of embedded words (vertical axis) as a function of phone length of the carrier word (horizontal axis). The figure is based on the CELEX lexicon for Dutch (nearly 322,000 unique lexical entries). The average number of embedded words increases about linearly with carrier word length. Embedded words occur like rain drops: their distribution appears close to the Poisson distribution, which explains that the standard deviation is about the square root of the mean.

A large number of embedded words may lead to a combinatorial explosion of word sequences. It is known, however, that not all embedded words play the same role in word comprehension; this role depends on location of the embedding and the

http://dx.doi.org/10.21437/Interspeech.2017-1048

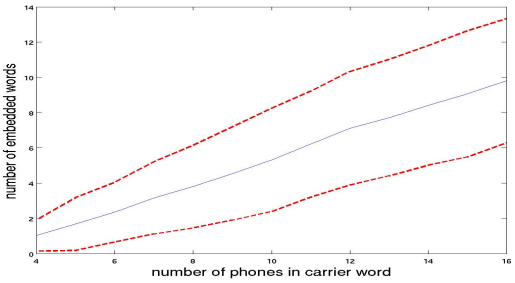

Figure 1: *The average number ($\pm 1\sigma$) of Dutch words (vertical axis) that are phonetically embedded in a Dutch carrier word as a function of phone length of the carrier word. In this computation, word frequency is not taken into account.*

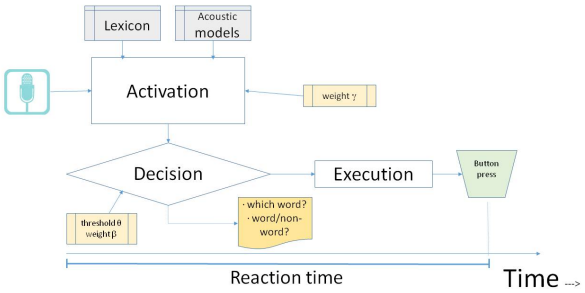

Figure 2: DIANA *consists of three interrelated components: (•) an Activation Component that takes speech as input; its output is a weighted lattice of hypotheses, evolving over time (•) a Decision Component, which outputs the recognized word/non-word item and an estimated RT (•) an Execution Component which models the time it takes from the mental decision until the eventual overt action (e.g., button press).*

listening condition [16]. In clear listening condition, embedded words that comprise a large proportion of the carrier word are activated regardless of their embedded position. Embedded words that comprise a small proportion of the carrier word are activated only when they are initial-embeddings. In BALDEY all stimuli were clearly pronounced in clean condition.

On the basis of these findings, we hypothesize that the density of competing word candidates changes over time during the unfolding of long words. In the survival risk analysis, such an effect should be reflected by a time-varying evolution of the regression between RT and word duration. In a complementary way, this effect should also be visible in the density and type of word decodings provided by DIANA as a function of time after onset. This paper provides a comparison of these methods.

2. DIANA

DIANA [12, 13, 14, 15] aims to simulate participants' behavior in experiments in spoken word comprehension. DIANA has accurately simulated lexicality judgments and corresponding reaction times for several different lexical decision experiments (Dutch: BALDEY, [9, 12, 13]; North-American: the Massive Auditory Lexical Decision data (MALD)[1] [14]), and behavioral data as in [17, 15]. In this paper we focus on the structure of DIANA's word sequence hypotheses.

DIANA consists of three components: an Activation Component, a Decision Component, and an Execution Component (cf. Figure 2). Activation and Decision operate in parallel; once the Decision Component has made a decision, the Execution component is initiated. Being a computational model of the cognitive processes involved in spoken word comprehension, DIANA does not simulate humanly effects such as waning attention or fatigue.

DIANA takes the acoustic speech signal as input, rather than a symbolic phone-like representation. This is in line with recent findings that indicate that phones cannot claim a plausible status as units in the cognitive process of speech recognition (see [18] in the context of perceptual learning; see [19] in the context of cortical activations).

The current implementation of DIANA can handle lexicons of about 40,000 entries. Each entry is accompanied by a prior probability, derived from a text corpus.

In the output of the Activation Component, the activation of a word sequence hypothesis is determined each 10 ms by combining acoustic bottom-up information and top-down information, based on the conditional prior probability of the words in the hypothesis:

$$\log P(\text{signal[0:t]}|\text{word}) + \lambda \log(P(\text{word}|\text{precontext})) \quad (1)$$

in which the parameter λ governs the balance between the bottom-up acoustic information (first term) and the top-down linguistic information (second term). $P(\text{word}|\text{precontext})$ combines a word entrance penalty ρ ($\rho \geq 0$, modeling a '0-gram') with an n-gram ($n \geq 1$). A low value of ρ will lead to many short word candidates; larger values of ρ yield longer words. The parameter ρ models the perceptual/cognitive bias to combine shorter words into a longer word (see, e.g., [4, 20]). Because this bias directly influences DIANA's word search at a low level, it is likely that it better describes the cognitive process of finding embedded words than a conventional n-gram LM would do, since an LM is trained on complete words in serial order, rather than on partially pronounced embedded words within carrier words.

The acoustic waveforms of the NN compounds in BALDEY were used to construct input stimuli for DIANA. For each 10 ms step from stimulus onset to offset, DIANA takes as input the gated signal (signal$[0 : t]$). For the first steps into the stimulus, evidently this output list only contains either very short words (denoted SW) or words that are only partially pronounced ('truncated' words, denoted W$\star$). When more acoustic evidence becomes available, sequences of SWs or Ws emerge, possibly followed by W$\star$. DIANA's lexicon consists of the 896 NN compounds and the 38,000 most frequent Dutch words.

3. Experimental set-up and data

Since there are many types of compounds with different cognitive processing routes [20], we limited the stimulus set to noun-noun (NN) compounds. We used the 896 NN compounds in BALDEY [9]. The BALDEY dataset contains behavioral data from a large-scale Dutch lexical decision experiment comprising 5,541 different stimuli from 20 participants, with in total over 110,000 RTs and lexicality judgments.

The 896 compounds were either real words or pseudo words. All pseudo-word compounds in BALDEY were the combination of two shorter pseudo words. Of all 17,786 responses, 8,915 and 8,871 responses were given to real and

[1] http://aphl.artsrn.ualberta.ca/?p=517

Table 1: *Results of the risk model*

test on significance of predictor β's				
	word resp.		pseudo-word resp.	
Predictor	statistic	p	statistic	p
(Intercept)	10.65	0.000	4.23	0.0000
FormFreqSc	3.34	0.014	1.54	0.4638
WordDurSc	5.32	0.000	3.60	0.0026
realword	4.57	0.000	7.37	0.0000
Kolmogorov Smirnov test on constancy				
	statistic	p	statistic	p
(Intercept)	1.336	0.0956	0.0127	0.0002
FormFreqSc	0.689	0.0290	0.0013	0.7114
WordDurSc	0.247	0.0000	0.0085	0.0000
realword	2.277	0.0306	0.0212	0.0300

pseudo-word stimuli, respectively. RTs below 100 ms were discarded, leaving 17,786 responses pertaining to NN compounds.

The average duration of all NN compound stimuli is 938 ms ($\sigma = 136$ ms); the average stimulus duration in BALDEY is 690 ms ($\sigma = 183$ ms).

The RTs pertaining to the NN compounds available in BALDEY form the basis for the risk analysis; the acoustic stimuli of the NN compounds form the input for DIANA.

4. Results

4.1. Risk analysis

We applied the following risk model on the RT data for the BALDEY NN compounds. The risk analysis model uses the pair `(RT, cause)` as dependent variable, in which the 'cause' is either a 'real word' or 'pseudo-word' judgment. As predictors, we used the conventional predictors 'form frequency', 'stress pattern', 'stimulus duration', and 'real/pseudo word'. Intercepts are included as random effect (called 'cluster') under subject.

$$\begin{aligned} &comp.risk(Event(RT, cause) \sim \\ &FormFreqSc + const(initial\_stress) + \\ &WordDurSc + realword + cluster(cluster), \\ &data = data, cause = 1, resample.iid = 1, \\ &n.sim = 5000, model = "additive") \end{aligned}$$

Table 1 presents the results of the risk model that was significantly better than models in which predictors were modified or deleted. The first and last 2 columns refer to real word responses and pseudo-word responses, respectively. The upper table presents the result in terms of regression coefficients. For word responses, intercept and the βs of all predictors are significantly different from 0; the word frequency (FormFreqSc) has the lowest significance. For pseudo-word responses, word frequency looses significance (which is not surprising since these responses are given to the pseudo word stimuli that are all allotted a low frequency). The lower panel shows the result of a Kolmogorov-Smirnov test which indicates whether the β's in the upper half of the table are constant or changing over time. The impact of word duration is clearly significantly different from a constant, which shows that the effect of word duration on RT changes during the unfolding of the stimulus.

Figure 3 displays the results in another way. Each subplot shows the dependency of the β of a predictor as a function of time (t). The six plots show this dependency for the intercept

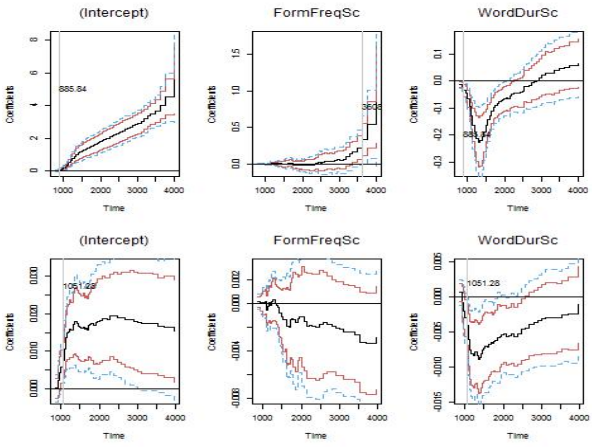

Figure 3: *Output of the risk model.*

(left two subplots), the form frequency (middle two subplots) and the word duration (right two subplots). The horizontal axis is the time (in ms) after stimulus onset. The black curve shows the mean, the red curves at either side mark the confidence range at $\pm 2\sigma$. The upper and lower row show separate risk analyzes for the 'real word' and 'pseudo word' responses, respectively. The somewhat noisy effects beyond 3500 ms are due to the small number of late RTs in the data.

The figure shows (in line with Table 1) a very significant effect of word duration on the β related to word duration: across the time interval from 1 to 3 s after onset, the effect first becomes negative, but increases again to become positive around 3 s after onset, showing that the pattern in the RT data changes during the unfolding of the stimulus. For pseudo word responses this 'early vs. late' effect comes later and is smaller in size. The risk analysis does not directly provide insight into these effects, but we will show below that the analyses by DIANA might provide a useful clue.

4.2. DIANA

Fig. 4 shows an example of the complexity that arises as DIANA decodes the Dutch compound 'kasteelhoeve' (kasteel *castle* + hoeve *farm*). Along the horizontal axis, the time t (in 10 ms steps) ranges from stimulus onset (0) to offset (840 ms). Activation is displayed along the vertical axis. Each trace shows the first-best hypothesis on the 84 gated signals signal$[0 : t]$. Each hypothesis may consist of either a truncated word (e.g., 'kasteelhoe' (orth.), /k ɑ s t e l h u / (phon.)) or a sequence of shorter words (SW^n), optionally followed by a truncated word (SW^n W$\star$). For the sake of clarity, the decoding result is only shown for a subset of the candidate sequences (truncated words are phonetically represented). The figure shows that until about 250 ms into the word all hypotheses have about the same activation score. After 250 ms, activations diverge, and later in the word a further ramification of candidates takes place. Interestingly, we observe a 'garden path' ('late revision') effect at around 600 ms and 750 ms after onset: hypotheses that seem to win in the beginning loose in the light of later acoustic evidence. These late-revision effects occur in 93% of the NN compounds in BALDEY.

Similar to Fig. 4, decoding data by DIANA were collected for each of the 896 noun-noun compounds, for a range of values of ρ, for each 10 ms gate step (t). The decoding results are summarized in Fig. 5, for real and pseudo compounds. The

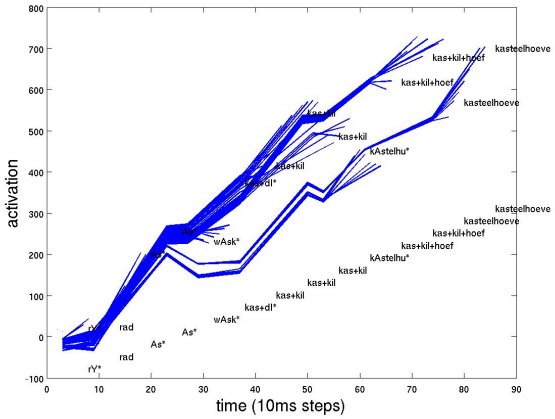

Figure 4: DIANA*'s decoding results for the compound 'kasteelhoeve'. Along the horizontal axis, the time t is displayed until which the speech signal is presented (input = signal$[0:t]$). Along the vertical axis, the activation is shown of the first-best word sequences matching the gated input.*

horizontal axis displays the word duration (relative: $0 =$ onset, $1 =$ offset); the vertical axis shows ρ. Each point (t, ρ) is labeled with the majority of types of DIANA decoding results for that (t, ρ)-combination.

These figures are best interpreted in terms of regime changes along horizontal lines, i.e., along the time axis. For high values of ρ, the majority of decoding results consist of single truncated words; only in case of a real compound, the compound pops up towards offset (Fig. 5 top-right corner). For lower values of ρ, DIANA shows a tendency to break up the decoding result into shorter chunks. At the beginning of the stimulus, this necessarily leads to a single short word or a truncated word; when more evidence becomes available, DIANA produces sequences of short words, possibly followed by an unfinished word. For very low values of ρ, the decoding of compounds nearly always leads to combinations of shorter words. This models the case in which a listener interprets a long stimulus in sequences of very short words. In general, five or six different 'decoding regimes' can be distinguished. The commonality between both plots are the similar regimes until around 2/3 of the stimulus, while the major difference between real and pseudo compounds occurs at around 2/3 of the stimulus duration (i.e., about 600ms after compound onset).

5. Discussion and conclusion

The psycholinguistic literature on the comprehension of compounds shows that there are several complex issues involved in their processing. In this paper, this complexity is shown in two complementary ways that are consistent with each other. The risk modeling suggests that, for real-word judgments, the RTs between 1000-1500 ms from word onset undergo a change in terms of a significantly different β value related to the stimulus duration. A similar effect, albeit a bit later and with a smaller size, holds for the pseudo-words responses. It has been shown that the influence of the word priors in lexical decision experiments is usually small ([15] and references therein); this is again supported by this risk analysis (Table 1).
DIANA's results provide a clue for explaining this 'early vs. late' effect, in terms of the density of revisions along the decoding path, especially in the first 2/3 into the word. Until about 2/3 of the stimulus, the variation in terms of decoding results and the garden path effects are more frequent than thereafter. This supports the dynamic behavior of word duration in Figure 3. It must be observed that the overt effects in the RT emerge at 1000 ms after onset, which is about 200-300 ms after the regime change detected in DIANA's decoding output.
In DIANA's output, the structure of the decoding (in terms of short words, sequences of short words, truncated words) is determined by the weighting between the acoustic model (bottom up) and the language model (top down) in Eq. 1, more specifically by the penalty ρ to enter a new word candidate. For the modeling of compounds, ρ is the most important parameter in DIANA. We argue that small changes in the parameter ρ may reflect subtle differences in the participant's interpretation of the instruction in a lexical decision experiment (e.g., "press the button as soon as you have heard a complete word" versus "press the button as soon as you think you hear a word").
In [15], an analysis was presented how to combine the information from data-oriented models (such as regression models, including the GAM) and process-oriented models such as DIANA. The results presented in this paper show that this is a fruitful path, but many details in line with [16] must be further investigated for a complete picture. It would be interesting to investigate human data on gated compound stimuli, but also challenging: in order to be insightful, such an experiment might turn out to be much larger than BALDEY.

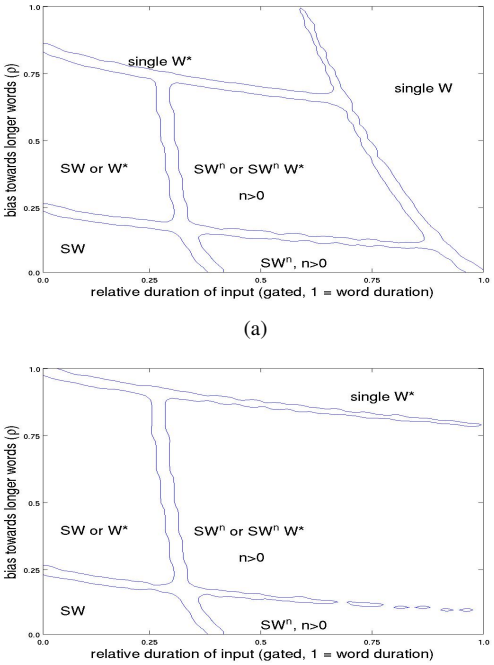

Figure 5: DIANA*'s decoding regimes in case of real compounds (upper plot) and pseudo compounds (lower plot).*

6. Acknowledgments

This work was funded by an ERC starting grant (284108) and an NWO VICI grant awarded to Mirjam Ernestus. The risk analysis is inspired by a risk modeling workshop by Paul Blanche, hosted by Harald R. Baayen, in Tübingen, Jan 2017.

7. References

[1] T. Brooks and D. Cid de Garcia, "Evidence for morphological composition in compound words using MEG," *Frontiers in Human Neuroscience*, vol. 9, 2015.

[2] D. Norris and J. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological Review*, vol. 115, pp. 357 – 395, 2008.

[3] G. Gaskell and W. D. Marslen-Wilson, "Ambiguity, Competition, and Blending in spoken word recognition," *Cognitive Science*, vol. 23, pp. 439–462, 1999.

[4] A. Cutler, *Native Listening: Language Experience and the Recognition of Spoken Words*. MIT Press, 2012.

[5] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, vol. 52, pp. 189–234, 1994.

[6] S. Mattys and J. Liss, "On building models of spoken-word recognition: When there is as much to learn from natural 'oddities' as artificial normality." *Perception and Psychophysics*, vol. 70, no. 7, pp. 1235–1242, 2008.

[7] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive Psychology*, vol. 18, pp. 1 – 86, 1986.

[8] O. Scharenborg, "Modeling the use of durational information in human spoken-word recognition," *Journal of the Acoustical Society of America*, vol. 127, pp. 3758 – 3770, 2010.

[9] M. Ernestus and A. Cutler, "BALDEY: A database of auditory lexical decisions," *Quarterly Journal of Experimental Psychology*, vol. Advance online publication, 2015.

[10] S. N. Wood, *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, 2006.

[11] B. Haller, G. Schmidt, and K. Ulm, "Applying competing risks regression models: an overview." *Lifetime Data Analysis*, vol. 19, no. 1, pp. 33–58, 2013.

[12] L. ten Bosch, L. Boves, and M. Ernestus, "Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task," in *Proceedings of Interspeech*, Lyon, France, 2013.

[13] ——, "Comparing reaction time sequences from human participants and computational models," in *Proceedings of Interspeech*, Singapore, 2014.

[14] ——, "DIANA: towards computational modeling reaction times in lexical decision in north american english," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

[15] ——, "Combining data-oriented and process-oriented approaches to modeling reaction time data," in *Proceedings of Interspeech*, San Francisco, USA, 2016.

[16] X. Zhang and A. Samuel, "The activation of embedded words in spoken word recognition," *Journal of Memory and Language*, 2015.

[17] I. Hanique, E. Aalders, and M. Ernestus, "How robust are exemplar effects in word comprehension?" *The Mental Lexicon*, vol. 8, no. 3, pp. 269–294, 2013.

[18] H. Mitterer, O. Scharenborg, and J. McQueen, "Phonological abstraction without phonemes in speech perception," *Cognition*, vol. 129, pp. 356 – 361, 2013.

[19] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[20] R. Fiorentino and D. Poeppel, "Compound words and structure in the lexicon," *Language and Cognitive Processes*, vol. 22(7), pp. 953–1000, 2007.