Towards a Unified Exemplar-Theoretic Model of Phonetic and Syntactic Phenomena

Hinrich Schütze (hs999@ifnlp.org)

Michael Walsh (michael.walsh@ims.uni-stuttgart.de)

Bernd Möbius (bernd.moebius@ims.uni-stuttgart.de)

Travis Wade (travis.wade@ims.uni-stuttgart.de)
Institute for NLP, University of Stuttgart, Germany

Abstract

In the last ten years or so, exemplar theory has enjoyed much growth in the field of phonetics. More recently, attempts have been made to apply exemplar theory to syntactic phenomena. Thus far, the issue of unifying phonetic and syntactic exemplar-theoretic models has not been addressed. This paper presents a single overarching exemplar-based model of constituent interactions across both linguistic domains which represents a significant first step towards a unified account of exemplar theory. Our simulations for one phonetic and two syntactic phenomena provide insights into how a unified account can be achieved. In addition, the phenomena we investigate shed light on the role of prototypes in exemplar theory and on whether exemplar clouds are defined by a fixed radius or by a fixed number of nearest neighbors.

Keywords: Exemplar theory; computational modelling; phonetics; syntax; acquisition; grammaticality; diachronic language change

1. Introduction

Exemplar-theoretic models are among the most successful in explaining human categorization (Nosofsky, 1986; Nosofsky & Zaki, 2002). There is also an increasing body of work applying exemplar-theoretic models to phonetic phenomena (e.g., Goldinger (1997), Johnson (1997)). Recent research in speech perception has provided considerable evidence indicating that the perception process is partly facilitated by accessing previously stored exemplars rich in phonetic detail. That is, speakers accumulate exemplars over time and compare input stimuli against them. Exemplars are categorized into clouds of memory traces with similar traces lying close to each other while dissimilar traces are more distant.

The appeal of exemplar models in phonetics is that they explain a number of phenomena that can pose problems for more abstractionist models. These phenomena include the detailed episodic memory of linguistic events that humans retain; the gradual change of categories in one speaker (as opposed to the speech community) in historical language change (Pierrehumbert, 2001); the plasticity of phonological categories (Norris, McQueen, & Cutler, 2003) and frequency effects in phonetics (Jurafsky, Bell, Gregory, & Raymond, 2001) and syntax (Bod, 2006; Bybee, 2006).

Our main contribution in this paper is that we present a unified model that explains phonetic as well as syntactic phenomena. The key innovation of the model is that it explicitly formalizes the relationship between exemplars on the *constituent* level and exemplars on what we call the *unit* level. Constituents are segments (e.g., consonants and vowels) in phonetics and words in syntax. Units are syllables in phonetics and phrases or sentences in syntax. Our simple hypothesis is that there is a competition between the submodel of the constituent level and the submodel of the unit level and that the unit submodel "wins" if the unit exemplar receives enough activation. A similar relationship between constituents and units holds in other models (e.g. Adaptive Resonance Theory (Grossberg, 2003)), but to our knowledge the model we present here is the first that explicitly models constituency in exemplar theory.

We will show that this simple competition model explains three different phenomena. The first phenomenon is variation in syllable duration, a phonetic phenomenon. The other two phenomena are syntactic: the grammaticalization of *going to* in English and the emergence of the notion of grammaticality in child language acquisition.

One of the important theoretical questions in exemplar theory concerns the status of prototypes. It has often been argued that a purely exemplar-theoretic account fails to explain many observations in human categorization (e.g., during early learning of a category, (Smith, Murray, & Minda, 1997)). The model proposed here is strictly exemplar-theoretic without any prototype component.

Finally, we address in this paper how exemplar clouds are formed. An exemplar cloud can be defined as either the k nearest neighbors around a stimulus or as all exemplars that have a distance of at most d from the stimulus, where k and d are parameters. We refer to these two types of exemplar cloud as *nearest-neighbor* and *radius-based.* We argue that for the two syntactic phenomena we consider, radius-based exemplar clouds are needed.

The paper is structured as follows. Section 2 introduces the unified exemplar-theoretic model. In Section 3, we use the unified model to explain variation in sylla-

	syllable duration	grammaticalization	grammaticality
stimuli	syllable to be produced	phrase (in perception)	phrase (in perception)
constituents	segments	words	words
constituent representation	acoustics, duration	word representation, left context, right context, tense	
similarity of constituents	sum of similarities of the components of the representation		
units	syllables	phrases	phrases
unit representation	sequence of constituents		
similarity of units	sum of similarities of the constituents of the units		
exemplar-based inference	duration of syllable	future tense	grammaticality of novel phrase

Table 1: Components of the unified exemplar-theoretic model.

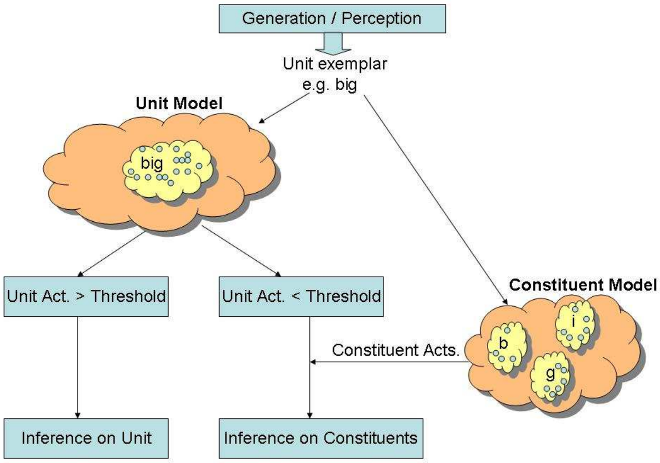

Figure 1: Architecture of the unified model. Example: The exemplar-theoretic inference process starts with the desire to articulate the word *big*. The exemplar cloud of *big* is computed in the unit (in this case: syllable) database. An exemplar cloud for each of the segments of *big* is also computed in the constituent (in this case: segment) database. The desired inference (in this case: duration) is then computed on the exemplar cloud(s) that were chosen based on greatest activation (unit vs. constituent).

ble duration in phonetics. In Section 4, we model the grammaticalization of *going to* as a future auxiliary in English. Section 5 applies the model to the acquisition of grammaticality. Section 6 discusses our experimental results, related work and future directions.

2. Exemplar-theoretic model

The architecture of the unified model is shown in Figure 1. The model has five components:

- A generation/perception component that generates (possibly underspecified) unit exemplars that serve as stimuli for the model. This component is either instantiated by a speaker different from the one that we are modeling (as when grammaticality judgments are modeled) or as the part of the cognitive system that determines which words or phrases are to be generated next. The unit exemplar *big* is an example for the latter case in the figure.
- An exemplar model on the unit level. The unit exemplar model retrieves all exemplars that are within a distance of at most d_u from the stimulus. If the activation the stimulus receives is above a threshold, then inference will be based on this unit exemplar cloud.
- An exemplar model on the constituent level. Operating in parallel with the unit level exemplar model, for each constituent of the stimulus, the constituent exemplar model retrieves all exemplars that are within a distance of at most d_c from that constituent. If the stimulus does not receive sufficient activation in the unit exemplar model, then inference is based on the resulting constituent exemplar clouds.
- An inference component. The inference component takes an exemplar cloud and infers a property of the stimulus from its nearest neighbors in the exemplar database. For example, the duration of a syllable is computed as the average duration of the members of its exemplar cloud.
- Parser and composer (not shown in the figure). Implicit in this model is a mechanism that parses a unit into its constituents and composes a sequence of constituents into a unit. This component is different for each of the three instantiations of the model. For example, the duration of a unit is equal to the sum of the durations of its constituents. The tense of a phrase of the form *going to walk* is the tense of the constituent word *going*.

Table 1 shows how the unified model is instantiated in the phonetic and in the two syntactic models. The following sections describe these instantiations in more detail.

Our methodology in this paper is to model the input data in a particular linguistic scenario (articulation, language change or language acquisition), present the model in Figure 1 with these input data, and then compare the predictions of the model with the outcome that was observed in the linguistic scenario.

3. Variation of syllable duration

In an exemplar model of speech production, exemplars serve as targets or plans of articulation. Schweitzer and Möbius (2004) note that speakers should have a significant number of exemplars for high frequency syllables, which would then act as a production target region, and a small or negligible number of exemplars for low frequency syllables. Consequently they argue that low frequency syllables would have to be computed online from exemplars of their constituents. They predicted, and observed, greater variation in duration for frequent syllables than for infrequent syllable.[1] The first simulation tests whether we can reproduce these experimental findings.

Stimuli. Stimuli were syllables of the form CVC where C was one of five consonants and V one of five vowels (for a total of 125 syllables). For each segment (phone) the acoustic properties are modeled as a randomly generated two-dimensional vector, and the duration value stored in a single dimension. The similarity of two segments or constituents was computed as the sum of the similarities of their acoustic vectors and their durations. For vector similarity, we employed the cosine, for duration similarity an exponential transformation of difference:

$$\mathrm{sim}(\vec{v}, \vec{w}) = \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}$$

$$\mathrm{sim}(x, y) = e^{-\alpha(|x-y|)}$$

where x and y are durations and $\alpha = 0.05$. α was chosen to give good sensitivity for typical lengths of consonants and vowels. Durations of syllables in the seed set were chosen to be 280 ms (but see Section 6), distributed in a ratio of 1:2:1 over the three constituents CVC. These numbers were chosen because 70 ms is a typical duration for a consonant and 140 ms is a typical duration for a vowel. The 125 syllables types were randomly assigned to either the frequent or the infrequent subclass.

Procedure. The unit exemplar database was seeded with 500 syllables. In all instantiations of the model, when a unit is added to the unit database, its constituents are simultaneously added to the constituent database.

We then ran 5000 iterations of a production-perception loop. Each iteration consists of randomly picking one of the 125 syllable types. If the type is rare, it is discarded with probability 0.99 and a new syllable type is generated. For the constituents of frequent syllables and infrequent syllables that survive the elimination step, acoustic vectors are generated (slightly perturbed, using uniform noise, from the canonical vector of a consonant or vowel to reflect variation in (planned) articulation). We then retrieve the syllable's and constituents' nearest neighbors in the unit and constituent databases respectively, within a fixed radius. If activation in the unit database is below the threshold θ_1 (i.e., there are fewer than θ_1 exemplars in the cloud), then the unit cloud is discarded, and the three neighborhoods in the constituent database are employed instead. The target duration of an exemplar is inferred to be the average duration of the members of its cloud. Finally, random noise proportional to the computed duration is added. The choice of the radius parameters and of θ_1 will be discussed below.

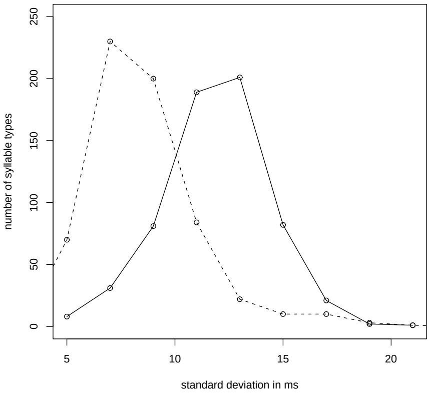

Figure 2: Experimental results for variation of syllable duration. Infrequent syllables (dashed line) have lower variability in duration than frequent syllables (solid line).

After the syllable with the inferred duration has been produced, it is added to the exemplar database. This part of the procedure models a production-perception loop, either on the individual or the community level: every produced exemplar becomes a perceived exemplar after its production.

The final phase of the procedure consists of probing the model, in an identical manner to the initial 5000 iterations, with 10 syllables of each of the 125 syllable types. The standard deviation for the syllable type is then computed on just this sample of 10 per syllable type. In this phase, syllables and their units are deleted after each probing to make sure that infrequent syllables do not change their status to frequent in the probing phase.

Results. Figure 2 is a cumulative histogram of 10

[1] Note that Schweitzer and Möbius (2004) found that z-scores of frequent syllable durations were more variable than z-scores of infrequent syllable durations. We interpret this here to mean that frequent syllables are more variable in duration than infrequent syllables. We are currently conducting further analysis of their data to confirm the validity of this interpretation.

runs of the above experiment, corresponding to 1250 standard deviations. The model successfully simulates the finding of Schweitzer and Möbius (2004): frequent syllables are more variable in duration than infrequent syllables. This result was significant ($p < 0.001$, Welch Two Sample t-test on 634 rare and 616 frequent syllables).

The difference in variation arises from the interaction between the two submodels. Frequent syllables have enough density, so that their duration is computed in the unit model, with noise added that is proportional to the length of the syllable. Infrequent syllables are compositions of constituents that are computed in the constituent model, each with independent noise. Therefore, the noise components often cancel out. Over many iterations of the production-perception loop, frequent syllables become more variable in duration whereas the variability of infrequent syllables does not change much.

4. Grammaticalization of *going to*

Starting in the 17th century, the construction *going to* was grammaticalized in its use as a form of future tense. We chose to model this phenomenon because it is often used as a prototypical example of the role frequency plays in language change.

One hypothesis is that this grammaticalization was caused by the temporary rise in frequency of phrases like *moving to do* with the connotation of intention and future (where *moving* is any motion verb) (Tabor, 1994; Bybee, 2006). Additional facts about the English of the 17th century (and today's English) are that *to go* is the most frequently used motion verb and that there are many more literal uses of motion verbs (motion to a location or to an object: *went to London*) than "verbal" uses like *running to meet*. We will show presently that based on these three assumptions, the unified model predicts the grammaticalization of *going to* as a future tense. We begin by motivating the representation of words in the unified model.

Representation of words. The similar syntactic behavior of two nouns like *cow* and *hen* is not directly apparent from their pronunciation or semantics. But an exemplar-theoretic account of syntactic behavior requires a similarity metric where *cow* and *hen* are similar. Building on the ideas described in (Schütze, 1995), we define left-context and right-context components of the representation of a given focus word, where the left (right) context consists of a probability distribution over all words that occur to the left (right) of the focus word and the dimensionality of the vector for each word is dependent on the number of distinct neighbors (left and right). For example, if we have experienced *take cow* twice and *drop cow* once, then the left context distribution of *cow* is $P(\text{take}) = 2/3, P(\text{drop}) = 1/3$. The similarity of two left context distributions can then be computed from the Jensen-Shannon divergence (which we again transform into a similarity using $\exp(-\alpha x)$, here: $\alpha = 5$):

$$0.5(D_{KL}(P||\frac{P+Q}{2}) + D_{KL}(Q||\frac{P+Q}{2}))$$

where P and Q are the probability distributions of the left (right) context of words 1 and 2, respectively, and D_{KL} is the Kullback-Leibler divergence (which we do not use as a distance measure because it is asymmetric and undefined if there is a single word that occurred in only one of the two left (right) contexts, giving rise to a 0 probability).

The intuition behind this representation of words is that we remember the typical left and right contexts of words. Two left (or right) contexts are similar to the extent that the distributions of words occurring in them are similar.

Future and motion are represented as two different four-dimensional vectors (as before, noise is added each time a tense or motion vector is generated to reflect slight contextual differences). Finally, the word itself is also represented as a four-dimensional vector. The similarity of two words is then computed as the sum of the similarities of the four components just enumerated: left context, right context, future/motion, and word.

Stimuli. In this simulation, five different constructions were presented to the model. We give an example for each: *going to fetch*, *going to Peter*, *walking to fetch*, *walking to Peter*, and *Peter fetch(es)*. Sentences of type *going to fetch* and *walking to fetch* are either generated as future sentences or as motion sentences. There were four moving verbs like *walking* in addition to *going*, five different non-moving verbs like *fetch* and five different nouns (objects or locations) like *Peter*. To model the three observations of historical English outlined above, *going* was as frequent as the other four moving verbs combined; 75% of *walking/going to fetch* sentences were generated with future, the rest with motion; and sentences of type *walking/going to Peter* were always generated with motion and twice as likely as *walking/going to fetch* sentences.

Procedure. 2000 sentences were generated according to the distribution described. Left and right context vectors for each word were computed for these 2000 sentences. The model was then presented with 30 sentences each of types *going to fetch*, *walking to fetch*, and *going to Peter*. If activation of the unit exemplar cloud was high enough, the prevalence of future tense was computed as the percentage of phrases in the unit exemplar cloud that were in future tense. Otherwise the prevalence was computed on the constituent exemplar cloud of the verb (*walking*, *going* etc).

Results. Figure 3 shows cumulative histograms for 10 runs. We assume a suitable competitive behavior between motion and future, so that only the more strongly

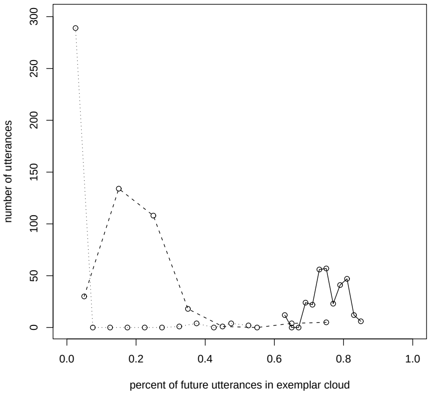

Figure 3: Experimental results for grammaticalization of *going to*. Histogram for strength of future tense in exemplar cloud for sentence types *going to fetch* (solid), *walking to fetch* (dashed) and *going to Peter* (dotted).

activated alternative survives. Thus a percentage of 60% would correspond to future, a percentage of 40% to motion.

In 99.3% of cases the future tense was not inferred for *going to Peter* sentences (future inference only occurred with activations in excess of 0.5, and 96.3% of the activations which were less than or equal to 0.5 were 0). For *walking to fetch* sentences the prevalence of future uses was consistently below 40%, for *going to fetch* consistently above 60%. Thus, the model correctly predicts the three key phenomena that occurred in the grammaticalization of *going to*: (i) *going to fetch* is grammaticalized as future tense; (ii) the other moving verbs are not grammaticalized and instead retain their original motion sense; and (iii) sentences of type *going to Peter* also retain their original motion sense.

The basic mechanism responsible for the simulation result is again the competition between the two levels. Sentences of type *going to fetch* have dense exemplar clouds due to their frequency and are processed on the unit level. Sentences of type *running to fetch* have sparse exemplar clouds due to their infrequency and are processed on the constituent level where there is no prevalence of future uses. Sentences of type *going to Peter* are not similar on the unit level to *going to fetch* because of the different left and right contexts of (proper) nouns like *Peter* and verbs like *fetch*.

5. Grammaticality judgments

In this section, we show that grammaticality judgments in the unified model can be formalized as activation of a sentence as a unit. The reasoning is that when a sentence does not give rise to enough activation as a unit, but is represented by an activation pattern of separate words, then it is perceived as ungrammatical.

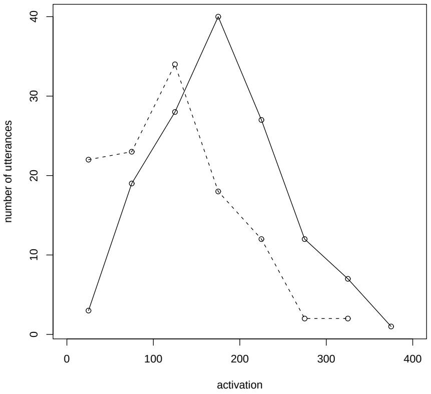

Figure 4: Experimental results for grammaticality judgments. Attested sentences (solid line) receive slightly higher activations than unattested grammatical sentences (dotted line). All 250 ungrammatical sentences in the 10 runs received an activation of 0 (not shown).

Stimuli. Using 5 different verbs and 5 different nouns, 25 sentence types of the form *I verb noun* (e.g, *I love coffee*) were generated and randomly assigned to the subclasses attested and unattested. In addition, 25 ungrammatical types of the form *I coffee love* were also generated. The same representation for words as in the previous experiment was used.

Procedure. In 1000 iterations, an "attested" grammatical sentence was generated and stored in the model. No ungrammatical and no unattested sentences were stored. An instance of each of the 25 grammatical and of the 25 ungrammatical sentences was then presented to the model.

Results. Figure 4 shows cumulative histograms for 10 runs. While unattested grammatical sentences receive slightly lower activation than attested sentences, they clearly are close to the distribution of grammatical sentences. In contrast, no ungrammatical sentence received any activation on the unit level. Thus, the model distinguishes grammatical (activation > 0) and ungrammatical sentences (activation $= 0$) with 100% accuracy.

The simulation succesfully models the acquisition of grammaticality because (i) attested and unattested sentences have very similar representations due to similar left and right contexts and (ii) ungrammatical sentences are dissimilar to grammatical sentences due to different left and right contexts. An example for the latter is that when comparing *I love coffee* with *I tea drink*, the left context of *love* (containing the subject *I*) is very different from the left context of *tea* (consisting of verbs like *love*, *drink* and *make*).

6. Discussion

We have presented a model that makes correct predictions for three linguistic phenomena. It is noteworthy that the model achieves this without prototypes or any explicit abstraction mechanism. Note in particular that Abbot-Smith and Tomasello (2006) express doubts in a recent paper that a pure exemplar-theoretic model can account for grammaticality judgments. Based on our simulations exemplar theory seems to provide an adequate account.

In our opinion, the experiments show conclusively that neighborhoods in exemplar theory must be radius-based as opposed to nearest-neighbor. In the case of grammaticality, even ungrammatical sentences have nearest neighbors (albeit neighbors that are far away). It is not clear how grammaticality judgments could be modeled with nearest-neighbor clouds. Similarly, the difference between the grammaticalization of *going to fetch* vs. non-grammaticalization of *walking to fetch* also requires a fixed-radius neighborhood. Previous arguments for nearest-neighbor clouds were based on difficulty of implementation (Pierrehumbert, 2001) and not on any fundamental reasons.

One challenge for exemplar theory is to explain how exemplars of constituents interact with exemplars of compositions of constituents into larger units. Segments and words on the one hand, and syllables and phrases on the other hand, each give rise to exemplar clouds at different levels. One of the key properties of language is the interaction of such units at different levels. The model proposed here is the first formal model of exemplar theory to address this issue.

The main deficiency of the work we have presented here is that we manually selected the parameters d_i (the radii of the exemplar neighborhoods) and the thresholds θ_i (the activation thresholds below which the constituent level is chosen). Obviously, the performance of the model depends on the values of these parameters. If the radius in the grammaticality model is too large, then even ungrammatical sentences will be judged grammatical (assuming a sufficiently small θ). However, we believe that these parameters can be estimated from the distribution of exemplars. For example, the distances of ungrammatical sentences from the nearest neighbor are much larger that that of grammatical sentences. We are currently exploring density estimation as one possible solution to this problem. In addition, although the syllable data here are simulated, parallel work with this model, employing the Schweitzer and Möbius (2004) corpus, has yielded z-score results in keeping with their findings.

Acknowledgments

This research was funded by the German Research Council (DFG, Grant SFB 732).

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*, 711–733.

Goldinger, S. D. (1997). Words and voices—perception and production in an episodic lexicon. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing.* San Diego: Academic Press.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423–445.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing.* San Diego: Academic Press.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure.* Amsterdam: Benjamins.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(5), 924–940.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure.* Amsterdam: Benjamins.

Schütze, H. (1995). Distributional part-of-speech tagging. In *Eacl 7* (pp. 141–148).

Schweitzer, A., & Möbius, B. (2004). Exemplar-based production of prosody: Evidence from segment and syllable durations. In *Proc. of the speech prosody conference* (pp. 459–462).

Smith, J. D., Murray, M. J., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 659-680.

Tabor, W. (1994). *Syntactic innovation: A connectionist model.* Unpublished doctoral dissertation, Stanford University.