ESSLLI 2007

19th European Summer School in Logic, Language and Information
August 6-17, 2007
Trinity College Dublin
Ireland

WORKSHOP PROCEEDINGS

ESSLLI is the Annual Summer School of FoLLI,
The Association for Logic, Language and Information
http://www.folli.org
Programme for "Exemplar Based Models of Language Acquisition and Use"

Workshop for ESSLLI 2007, Trinity College Dublin, 13-17 August 2007

The workshop runs in the late morning session, each day of the second week of ESSLLI, from Monday 13th to Friday 17th. Each session lasts for 90 minutes, comprising two talks; talks are given 45 minutes, including discussion time, unless otherwise stated.

Monday 13th

- "Introduction to Exemplar-Based Models of Language Acquisition and Use", Rens Bod & Dave Cochran (30 min)
- "Effects of Word-Chunk Frequency on Language Acquisition and Use", Invited Talk, Morten Christiansen (60 min)

Tuesday 14th

- "The Variability of Compound Stress in English: Towards an Exemplar-Based Alternative to the Compound Stress Rule", Sabine Lappe & Ingo Plag
- "Accounting for Phonetic and Syntactic Phenomena in a Multi-Level Competitive Interaction Model", Hinrich Schütze, Michael Walsh, Travis Wade & Bernd Möbius

Wednesday 15th

- "Modelling the steps of early syntax acquisition", Jacqueline van Kampen & Remko Scha
- "Exemplars in Syntax: Evidence from Priming", Neal Snider

Thursday 16th

- "HPSG-DOP: Towards Exemplar-based HPSG", Doug Arnold and Evita Linardaki
- "Analogical Modeling of Language: An Update", David Eddington & Deryle Lonsdale

Fridays 17th

- "From Exemplar Theory to Population Coding and Back. An Ideal Observer Approach", Laurent Bonnasse-Gahot & Jean-Pierre Nadal
- General Discussion Session and Concluding Remarks, chaired by Rens Bod & Dave Cochran
Introduction to Exemplar-Based Models of Language Acquisition and Use

Rens Bod and David Cochran
School of Computer Science
University of St Andrews
St Andrews, Scotland, UK
{rb,davec}@cs.st-and.ac.uk

1 The notion of exemplar

Exemplar-based models have become part and parcel of linguistics: they are used in phonetics and phonology (Johnson 1997; Pierrehumbert 2001), in morphology (Skousen 1989; Krott et al. 2002), syntax (Bod 1998; Tomasello 2003; Reali and Christiansen, in press) and semantics (Batali 2002). The appeal of exemplar models is that they can explain phenomena that are problematic for rule-based models. These phenomena include frequency effects in comprehension and prediction, gradual language change, and the dynamics of language learning. Exemplar models aim to capture the detailed episodic memory of linguistic events that humans retain, by storing exemplars over time and comparing new input against them.

But what exactly are exemplars? A quick look at the literature suggests that there is no agreement on what an exemplar is and what it does. One goal of this workshop is to clarify the various notions of exemplar and to investigate the possibility of a unifying framework. Following Hay and Bresnan (2006), there are at least two strands of exemplar-based models that were developed rather independently of one another, and to which they refer as “phonetic exemplar theory” and “syntactic exemplar theory”.

2 Phonetic Exemplar Theory

Exemplars were introduced as a psychological model for perception and categorization by Hintzman (1986), Nosofsky (1986) and others. These exemplar models have been investigated in speech perception and production where the representation of lexical items are conceptualized as underlying, abstract forms (Johnson 1997). Every time we encounter a particular word, we store the phonetic memory of that word together with non-linguistic detail. When a new word is encountered, it is classified according to its similarity to the exemplars already stored. Its similarity to any single stored exemplar can be computed as its ‘distance’ from the exemplar in the parameter space. The most probable labelling given the labelling of the exemplars in the neighborhood is computed on the basis of stored exemplars. Pierrehumbert (2001) views exemplar theory as providing a way to formalize the detailed phonetic knowledge that speakers have about the categories of their language. The acquisition of this knowledge can be understood in terms of the acquisition of a large number of memory traces of experiences.

This exemplar-based view appears in many different frameworks. It has been used by Skousen (1989), Krott et al. (2002) and Daelemans and van den Bosch (2005) to model analogical processes in various subfields. The differences between the frameworks lie in the definitions of similarity and distance metrics and the functions that are maximized. But the common underlying idea is that all previous linguistic experiences are stored as exemplars and linguistic behavior is predicted by the most similar stored exemplars.

3 Syntactic Exemplar Theory

According to the exemplar-based conception in syntax, there are no explicit rules of grammar. Instead, a ‘grammar’ arises out of analogical generalizations over stored chunks of previous language experiences. Language experiences can
be sequences of words, phrases or entire sentences. Syntactic exemplar models are associated with Construction Grammar (Fillmore et al. 1988; Goldberg 2006), Cognitive Grammar (Langacker 1998) and Data-Oriented Parsing (DOP) (Scha 1990; Bod 1998). The syntactic exemplar model that has been worked out in most formal detail is DOP. In this model, exemplars correspond to syntactic structures of observed utterances. New utterances can be produced or analyzed by combining largest possible and most frequent chunks from stored exemplars. In this way, DOP offers a full exemplar-based account of grammatical productivity. Language acquisition can be understood by assigning arbitrary categories to phrases which are next selected on their usefulness in processing new input (Bod 2006, 2007).

Both DOP and Construction Grammar combine general, abstract knowledge and specific, episodic memory into one model, similar to Abbot-Smith and Tomasello (2006) who opt for a hybrid model comprising both abstractions and the retention of exemplars of which those abstractions are composed. The work by Batali (2002), who investigates the emergence of semantic structures in language acquisition and evolution, can also be viewed in these terms.

4 Towards Unified Exemplar Theory?

There are substantial differences between phonetic and syntactic exemplars. While phonetic exemplar models mainly deal with the classification problem of how to store new input according to its similarity to the exemplars already stored, syntactic exemplar models focus on the composition problem of how new input is constructed out of interacting constituents from previous exemplars which are compositionally built into larger units. Since both classification and composition are part of language acquisition and processing, there is a quest for unification of the two types of models.

In the current absence of such a unification, we may describe exemplar theory loosely as follows: Exemplar theory as applied to language is a theory that involves storage of linguistic experiences and that allows for production and perception as analogical generalizations over the stored memories.

The papers in this workshop describe a variety of instantiations of this idea and some explore a far-ranging unification of phonetic and syntactic exemplar models. We let the papers speak for themselves and hope that you will enjoy the workshop.

Acknowledgments

Many thanks to Nick Chater, Alexander Clark, Walter Daelemans, Tecumseh Fitch, Susanne Gahl, Janet Pierrehumbert, David Tugwell, Antal van den Bosch, Menno van Zaanen and Jelle Zuidema for being part of the program committee.

References


Reali, F. and M. Christiansen, in press. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*.


Effects of Word-Chunk Frequency on Language Acquisition and Use

Morten H. Christiansen
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
mhc27@cornell.edu

Over the past two decades, a growing bulk of work has begun to stress the importance of linguistic experience in explaining the acquisition and use of language across a variety of disciplines, ranging from connectionist modeling (e.g., Christiansen & Chater, 2001) to natural language processing (e.g., Bod, 1998) to construction grammar (e.g., Croft, 2001) to developmental psycholinguistics (e.g., Tomasello, 2003). The timing of this workshop is very apt, as there currently seems to be a confluence of many of these approaches to language. Therefore begin my talk by pointing to areas of common ground; in particular, the importance of experience with specific exemplars.

A key issue for exemplar-based approaches in general is what counts as an exemplar. In this talk, I discuss research in which several different notions of exemplars are used—including sequences of specific word tokens, combinations of word tokens and lexical categories, as well as combination of lexical categories—to provide evidence in support of an experience-based account of language acquisition and use.

On the acquisition side, I show how indirect statistical information can be used to overcome the so-called poverty of the stimulus in learning how to form questions involving auxiliary fronting in polar interrogatives (e.g., Is the man who is smoking tall). I report on corpus analyses of child-directed speech indicating that simple learning devices can utilize information inherent in the co-occurrence of pairs (bigrams) and triples (trigrams) of consecutive words to produce grammatical questions (Reali & Christiansen, 2005).

In terms of language use, I then discuss how sensitivity such word-chunk exemplars (i.e., combinations of specific words and/or lexical categories) in adulthood can facilitate language processing, focusing on relative clauses. Using a series of large-scale corpus analyses and self-paced reading experiments (Reali & Christiansen, in press a, b), I show how on-line processing of relative clauses is affected by fine-grained distributional information.

Finally, I conclude by arguing that exemplar-based approaches to language may be best pursued within a cross-disciplinary approach involving multiple-cue integration.

1 Credits

My talk is to a large extent based on collaborative work with Florencia Reali. I’m grateful to have had the opportunity to work with her during her graduate studies at Cornell University.

References


The Variability of Compound Stress in English: 
Towards an Exemplar-Based Alternative to the Compound Stress Rule

Sabine Lappe
University of Siegen
lappe@anglistik.uni-siegen.de

Ingo Plag
University of Siegen
plag@anglistik.uni-siegen.de

Abstract

Recent research has shown that the Compound Stress Rule cannot adequately handle the variable stress behaviour of noun-noun constructs in English (e.g. Plag, 2006, Plag et al., 2006a, 2006b). In this paper we present an analysis of compound stress assignment in exemplar bases models, using data from two corpora, CELEX (Baayen et al., 1995) and the Boston University Radio Speech Corpus (Ostendorf et al., 1996). The data were coded according to a number of semantic and structural criteria taken from the literature on compound stress. Two different algorithms are tested, TiMBL 5.1 (Daelemans et al., 2004) and AM::Parallel (Skousen et al., 2004), and it turns out that both analogical algorithms are superior to previous, categorical approaches.

Credits

This research was funded by the Deutsche Forschungsgemeinschaft (Grant PL 151 / 5-1). Special thanks go to Gero Kunter for his splendid management of the CELEX and BURSC data, and to Maria Braun for helping us code the data.

1 Introduction

The Compound Stress Rule (Chomsky & Halle, 1968) is one of the most wellknown stress rules in English. It states that in English noun-noun compounds, the left-hand constituent is more prominent than the right-hand constituent. However, it is also wellknown that not all English noun-noun compounds abide to the Compound Stress Rule. Examples of both left-stressed and right-stressed noun-noun compounds are provided in (1). The most prominent syllable is marked by an acute accent.

\[ \text{Ópera glasses steel bridge} \]
\[ \text{wátch-maker morning páper} \]
\[ \text{clássroom silk tíe} \]
\[ \text{Oxford Street Madison Ávenue} \]

Rightward stress is far from exceptional, and the nature of the observable variability is still rather unclear. Recent investigations (e.g. Plag, 2006, Plag et al., 2006a, 2006b) have shown, however, that categorical approaches (such as the Compound Stress Rule) are unsuccessful in making correct predictions about compound stress assignment. This paper will present an analysis of the variation in compound stress assignment in exemplar-based models. We will use two current exemplar-based algorithms, TiMBL (Daelemans et al., 2004) and AM (Skousen et al., 2004), to investigate whether an exemplar-based approach to compound stress is empirically more adequate than a categorical or probabilistic model. As data we will use the data from the two Plag et al. (2006a, 2006b) studies. These data comprise all noun-noun compounds extracted from the Boston University Radio Speech Corpus (BURSC, Ostendorf et al., 1996) and all compounds extracted from the CELEX lexical database (Baayen et al., 1995).

The paper is structured as follows. In section 2 we will set the stage, introducing our theoretical assumptions (section 2.1), the BURSC and CELEX data and coding (section 2.2), and the methodological principles that guided our TiMBL and AM experiments (section 2.3). Sections 3 and 4 will then report our findings. The paper ends with a conclusion and outlook onto future research.
2 Setting the Stage

2.1 English compound stress

Whereas it is still general common ground that the Compound Stress Rule is the major predictor of compound stress, there are three types of approaches to deal with the observable variation in compound stress assignment, which we may label syntactic, semantic, and analogical. Right stress has been claimed to be an effect of the syntactic, i.e. phrasal nature of a construction, whereas left stress is associated with lexical, i.e. morphological, items. A recent proponent of this structure-based approach is Giegerich (2004), who claims that argument-head compounds such as bookseller are morphological entities and thus generally left-stressed, while modifier-head compounds are generally stressed on the right. Apparent exceptions, such as left-stressed opéra glasses are the result of the lexicalization of an originally phrasal structure.

Other people (e.g. Sampson, 1980, Fudge, 1984, Ladd, 1984, Liberman and Sproat, 1992, Olsen, 2000, 2001, Spencer, 2003) have claimed that right stress is triggered by specific semantic relations between left and right constituents (such as, for example, material or locative relations as in steel bridge or Boston hárbor). Finally, exceptional right stress has been explained as the effect of analogy (e.g. Schmerling, 1971, Liberman and Sproat, 1992), with compounds having the same right or left constituent sharing the same stress pattern. Standard examples of the latter phenomenon are street names, which are stressed on the left if the right constituent is street, but right-stressed if the right constituent is avenue or lane (as in Oxford Street vs. Oxford Avenue, Oxford Láne).

These different approaches have recently been tested against large amounts of empirical data by Plag et al. (2006a, 2006b), who investigated the phonetic properties and the determinants of compound stress assignment in several thousands of compounds in two corpora, CELEX and BURSC. The CELEX database (Baayen et al., 1995) contains mostly dictionary, i.e. lexicalised, data, while the Boston Radio Speech Corpus (BURSC, Ostendorf et al., 1996), contains audio recordings of radio news texts. In their studies Plag et al. found a number of new interesting insights about the nature of compound stress, the most relevant of which are:

- Neither the Compound Stress Rule nor any of the syntactic or semantic or analogical factors proposed in the literature can adequately explain compound stress in a categorical fashion.
- Statistical analysis reveals a relatively large amount of unexplained variation. Variation is found among compound types as well as among tokens of the same type.
- In spite of the fact that compounds from CELEX and BURSC exhibit striking differences with respect to their status of lexicalisation, the two corpora yield strikingly similar, almost parallel findings.

For the present study in exemplar-based modelling we used the data from the Plag et al. studies, as well as their coding of the data, to which we now turn.

2.2 The coding

In order to test the effects of argument structure, semantics and analogy, Plag et al. first extracted all NN structures from BURSC and CELEX. In what follows we will use the term ‘compounds’ as a convenient label to refer to these structures. We thus remain deliberately agnostic with respect to the question of whether or not some of these structures should be attributed a phrasal status. It should be emphasised, however, that all our compounds are of a kind to which is attributed word status, not phrasal status, in the general descriptive literature. The total number of NN constructions extracted from BURSC is 4410 tokens, which are distributed among 2476 different types. The total number of NN compounds extracted from CELEX is 4491 (types).

In the present study we used only types, not tokens. Furthermore, we used subsets of the two corpora which comprise those compounds that have a constituent family, i.e. for which there are other compounds that share the same right or left constituent. The rationale behind this choice was that we wanted to make sure that all sources of information were available for all compounds.

---

1 While this is straightforward for the lexical database CELEX, all texts from the BURSC had to be manually annotated for all sequences consisting of two (and only two) adjacent nouns, one of which, or which together, functioned as the head of a noun phrase. From this set proper names such as Barney Frank and those with an appositive modifier, such as Governor Dukakis were eliminated. The exclusion of these two types of structures was based on two considerations. First, we would expect these to show consistent rightward stress, second we know of no claims that these structures would be regarded by anyone as compounds.
For reasons of consistency, we used the same subsets in all experiments. For BURSC the total number of compounds in our subset is 722. The corresponding number for CELEX is 2643. All compounds were coded in terms of

- the orthographic representation of their left and right constituents
- the structural and semantic features held to be responsible for stress assignment in the literature
- the stress category (left or right)

For each compound the structural and semantic features were coded independently by two raters. The coding categories are given in (2) – (4). We broadly distinguish between three types of features: argument structure, semantic categories, and semantic relations. In what follows we will use the terms 'argument structure', 'semantic categories', and 'semantic relations' as convenient labels to refer to these sets of features.

(2) Argument Structure

<table>
<thead>
<tr>
<th>Constituent Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>argument-head</td>
<td>computer maker</td>
</tr>
<tr>
<td>modifier-head</td>
<td>truck accident</td>
</tr>
<tr>
<td>morphology of head: -er</td>
<td>computer maker</td>
</tr>
<tr>
<td>morphology of head: -ing</td>
<td>arts funding</td>
</tr>
<tr>
<td>morphology of head: -ion</td>
<td>habitat acquisition</td>
</tr>
<tr>
<td>conversion</td>
<td>budget cut</td>
</tr>
</tbody>
</table>

(3) Semantic property of constituent or compound

N1 refers to a period or point in time
example: day care

N2 is a geographical term
example: bay area

N2 is a type of thoroughfare
example: state road

The compound is a proper noun
example: Harvard University

N1 is a proper noun
Mapplethorpe controversy

(4) Semantic relation between the constituents of the compound

<table>
<thead>
<tr>
<th>Relation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2 CAUSES</td>
<td>N1 retirement age</td>
</tr>
<tr>
<td>N1 CAUSES</td>
<td>N2 drug war</td>
</tr>
<tr>
<td>N2 HAS</td>
<td>N1 school district</td>
</tr>
<tr>
<td>N1 HAS</td>
<td>N2 state inspector</td>
</tr>
<tr>
<td>N2 MAKES</td>
<td>N1 computer company</td>
</tr>
<tr>
<td>N1 MAKES</td>
<td>N2 university research</td>
</tr>
<tr>
<td>N2 IS MADE OF</td>
<td>N1 paper drum</td>
</tr>
<tr>
<td>N2 USES</td>
<td>N1 biotech industry</td>
</tr>
<tr>
<td>N1 USES</td>
<td>N2 police effort</td>
</tr>
<tr>
<td>N1 IS</td>
<td>N2 jail facility</td>
</tr>
<tr>
<td>N1 IS LIKE</td>
<td>N2 crime wave</td>
</tr>
<tr>
<td>N2 FOR</td>
<td>N1 consumer advocate</td>
</tr>
<tr>
<td>N2 ABOUT</td>
<td>N1 health law</td>
</tr>
<tr>
<td>N2 LOCATED at/..</td>
<td>N1 neighborhood school</td>
</tr>
<tr>
<td>N1 LOCATED at/..</td>
<td>N2 school district</td>
</tr>
<tr>
<td>N2 DURING</td>
<td>N1 lifetime</td>
</tr>
<tr>
<td>N2 NAMED AFTER</td>
<td>N1 Mapplethorpe show</td>
</tr>
</tbody>
</table>

Due to the well-known fact that many compounds are ambiguous and can be interpreted as belonging to more than one of the above semantic categories, each of the semantic categories had to be coded individually for each compound as a binary category (with ‘yes’ and ‘no’ as values). In addition to the categories in (2) – (4), the data were coded according to left and right constituents. These were given in their orthographic form. A sample of a coded item is given in (5).

(5) The coding of clinic worker (BURSC)

<table>
<thead>
<tr>
<th>Constituent</th>
<th>Left</th>
<th>Clinic</th>
<th>Worker</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>constituents</td>
<td>left</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>argument structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>morphology of N1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>semantic categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>semantic relations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stress</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

At this point a note is in order with regard to the coding of the target category, stress. For CELEX we relied on the classification of stress information as given in the corpus. For BURSC, the classification into left and right stress was based on the items' acoustic cues, using the algorithm proposed in Kunter & Plag (2007). The basis of this classification comprises both measurement and perception data on the acoustic correlates of stress: pitch (f0), intensity, duration, and jitter (as a correlate of creaky voice) in the right constituent. Kunter & Plag (2007) and Plag et al. (2006a) have shown that this model is highly reliable in predicting listeners' perceptions of prominence relationships in compounds from the BURSC corpus.

Nevertheless, it is important to note that our classification of stress in BURSC and CELEX is
subject to two sources of error: First, the classification ignores the within-speaker, across-speaker and within-type (i.e. token) variability that actually exists in compound stress assignment (Kunter, 2007). In CELEX stress is treated as an invariant property of the compound lemmata, so that we have no information about how robustly left- or right-stressed a particular item is. In BURSC, we classified a given type as left- or right-stressed if the majority of tokens of that type had left or right stress. Second, the automatic classification of stresses according to acoustic cues inevitably generates some error. In spite of these problems, however, it turns out that the analysis of the compound data from the two corpora produces very robust findings with respect to the determinants of compound stress. Thus, despite all differences between the two corpora, the study based on BURSC (Plag et al., 2006a) and the study based on CELEX (Plag et al., 2006b) produced strikingly similar results. The same holds for the experiments in the present paper. We thus have good reason to believe that the only influence that classification errors may have on our study lies in slightly lower rates of predictive accuracy than we may expect without these additional sources of error.

2.3 The models -- TiMBL and AM

In sections 3 and 4 we will present the results of a series of tests in which we examine in how far two computational implementations of exemplar-based models, TiMBL (version 5.1, Daelemans et al., 2004) and AM::Parallel (Skousen et al., 2004), are able to predict the actual distribution of left and right stresses in BURSC and CELEX. Both algorithms classify new items on-line by comparing a new test item with similar items that are stored in memory.

TiMBL is a k-Nearest Neighbour (k-NN) system that encompasses several different implementations of memory-based learning techniques (cf. Daelemans et al., 2004 for details). The classification of a new item is extrapolated from similar exemplars that are explicitly stored in memory (the item's nearest neighbours), via a majority vote (which may optionally be weighted according to the distance between a given neighbour and the test item). Nearest neighbours are selected from a distance space k. The experimenter can manipulate k, so that she has some control over how narrow this space should be for a given experiment. Also for the computation of similarity, the experimenter can choose from a variety of different similarity measures, which conceptually fall into two different classes. In one class of measures, similarity is computed in terms of the simple number of matching values for all features given in the new input (overlap metric). Alternatively, the degree to which matches between input features and features of stored exemplars are relevant for the computation of similarity is influenced by feature weights (Information-gain feature weighting and others), or by weights which are able to distinguish between different values for a single feature (MVDM, Jeffrey divergence metric). Again, there are different ways in which feature weights may be computed. Crucially, however, the features weights are computed on the basis of the whole dataset that the system is given as training data. As a consequence, if feature weights are used in an experiment, they will be the same for every new input that is to be classified. Thus, for example, if the algorithm has found that in the training set the right constituent is more informative than the syntactic relation, it will assign to the right constituent of every exemplar in memory a higher weight value than to the syntactic relation. For every new input, similarity will thus be computed using the same feature weights.

AM treats feature weights differently. In AM, the relevance of features for the classification of a given item is answered for every single new input on an individual basis (cf. Skousen, 1989, 2002a, b for details). The set of exemplars that is relevant for classification of a given input is termed the exemplar’s analogical set. For every input, the algorithm checks all conceivable combinations of features (termed contexts) and determines in how far the set of exemplars in memory that match that combination behave in a homogeneous way with respect to the target category (i.e., in our case, stress assignment). Only homogeneous contexts will then be taken into account for the analogical set (cf. Skousen, 1989, 2002a, 2002b for details). In this process, also contexts which are more general and, hence, less similar to the context of the test item may be taken into account. Apart from the huge processing demands that this kind of procedure entails, an interesting difference between TiMBL and AM therefore lies in the degree to which different features may play a different role for different inputs. Here AM seems to be more flexible.

In all experiments in the present study, we tested the corpus on itself. That means that every item in the corpus was classified on the basis of all other items present in the same corpus. Both
TiMBL and AM provide parameter settings that can be used to implement this kind of experimental setup. In TiMBL the relevant procedure is the leave-one-out procedure. In AM, we used the same data set for both training and testing and set the parameters in such a way as to ensure that those items in the training set that are identical to those in the training set are excluded during classification.

In our experiments we focus on three important aspects:

- symbolic, rule-based models vs. TiMBL, AM: Which models are empirically more adequate, measured in terms of predictive accuracy?
- abstract vs. non-abstract features in TiMBL, AM: How abstract do the features have to be for a successful computation of compound stress?
- nearest neighbours vs. analogical set: Do the differences between TiMBL and AM in which exemplars they consider for classification result in different predictive accuracies?

The first aspect concerns the question of whether TiMBL and AM are empirically more accurate than categorical, rule-based models. To this end, we will compare the two models' predictive accuracies with the predictive accuracies that the pertinent rule-based models reach for the data in BURSC and CELEX.

The second aspect concerns the question of which features are relevant for the representation of generalisations about compound stress. The coding of data in Plag et al. (2006a, 2006b) gives us the opportunity to compare three different claims about the nature of linguistic representation, along which recent theories are divided. On the one hand traditional accounts have claimed that stress is computed on the basis of quite abstract information about the semantic, syntactic, morphological, or semantosyntactic status of a compound. We will use the term 'abstract features' from now on as a descriptive label to refer to the pertinent features. On the other hand, however, recent studies have shown that the level of representation which is relevant for the computation of stress may not be all that abstract. This position is found, for example, in occasional remarks in the literature about analogical effects that occur with compounds with specific left or right constituents or, more radically, in exemplar-based theories, where analogical effects with previously encountered, 'non-abstract' representations of stored exemplars is held to be the rule, rather than the exception. Whereas work on other aspects of derivational morphology and compounding has produced growing numbers of evidence for this position (e.g. Gagné, 2001, Krott et al., 2002, Chapman & Skousen, 2005), this view has never been tested for English compound stress. The BURSC and CELEX data, which have been coded both in terms of the pertinent abstract as well as in terms of the 'non-abstract' lexical representation of the left and the right constituent, provide an ideal testing ground to look at abstractness of representation.

We will do this with the help of three different experimental series. In series 1 we test the simple and, admittedly, overly simplistic hypothesis that it is either argument structure, semantic categories, or semantic relations that can predict compound stress. The computational model is thus provided with only one set of the pertinent features. In the second experimental series we test in how TiMBL and AM are successful in predicting stress if fed with the most informative combination of abstract features conceivable. Finally, in series 3 we feed TiMBL and AM only with the 'non-abstract' features, the left and right constituents of each compound.

3 Modelling Compound Stress in BURSC

3.1 The data, or: the demise of the rule-based approach

Table 1 provides an overview of the actual distribution of stresses in the corpus (N = 722).

<table>
<thead>
<tr>
<th>left stress</th>
<th>right stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>358</td>
<td>364</td>
</tr>
<tr>
<td>49.58%</td>
<td>50.42%</td>
</tr>
</tbody>
</table>

Table 1. Distribution of stresses in BURSC.

The distribution in table 1 shows that the prediction of left stress as expected by the Compound Stress Rule does not correspond to the data. We will see in the analyses below that there is a huge amount of variation in the data, no matter according to which of the predictor categories we subdivide the data. As a consequence of that variation, not only the Compound Stress Rule, but also other traditional rule-based approaches to compound stress fail to account for the data. This has been shown in detail in Plag et al.'s BURSC.
study (2006a), whose main findings we will briefly summarise below.

Plag et al. test the predictive accuracy of approaches using argument structure (as proposed in Giegerich, 2004), semantic categories, and semantic relations (as proposed, e.g., in Fudge, 1984: 144ff., Liberman & Sproat, 1992, Zwicky, 1986), as predictors of compound stress. The central finding that emerged is that none of these three sets of predictors can adequately account for the variation that we find in the BURSC data. This is true for each set in isolation (argument structure, semantic categories, semantic relations) as well as for a combination of features from the three sets. This generalisation is independent of whether the features are conceptualised as predictors in a categorical rule model or in a logistic regression model. For reasons of space, we will limit the present discussion to rule-based approaches.

Table 2 provides an overall summary of the predictive accuracies for rule-based models that are reported in Plag et al. (2006a). We distinguish here between overall predictive accuracies and predictive accuracies for right and left stresses.

<table>
<thead>
<tr>
<th>Predictive Accuracies</th>
<th>Overall</th>
<th>For Left Stress</th>
<th>For Right Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argument Structure</td>
<td>53.8%</td>
<td>18.0%</td>
<td>85.3%</td>
</tr>
<tr>
<td>Semantic Relations</td>
<td>54.5%</td>
<td>30.3%</td>
<td>76.8%</td>
</tr>
</tbody>
</table>

Table 2. Predictive accuracies of categorical rules for BURSC (from Plag et al. 2006a).

It is important to note that the two approaches cannot be compared directly because, due to methodological constraints, the figures could not be computed from exactly the same set of the data (for the argument structure approach: $N = 4091$, for the semantic approach: $N = 2027$). However, in terms of the general distribution of all relevant predictor categories, the two subsets behave almost identically. The same is true for the subset of 722 items used in the present study.

Neither of the two approaches reaches an overall predictive accuracy that is significantly beyond a prediction by chance. If we look at predictive accuracies for left and right stresses in isolation, we see that both approaches are good at predicting right stress, but very poor in their predictions of left stress. This does not come as a surprise, given that both approaches have come into existence as an attempt to explain why there are so many exceptions to the Compound Stress Rule, i.e. why there are so many right-stressed compounds in English. What the table shows, however, is that they go too far in their predictions, grossly overpredicting right stress. Note that this also means that neither the syntactic nor the semantic approach can be saved if we combine them with the Compound Stress Rule. They would still predict right stress for the majority of items while their left-stress predictions converge with the predictions made by the Compound Stress Rule.

The question of the empirical accuracy of rule-based models is, however, made more complex than the findings in table 2 may suggest. Thus, Giegerich (2004) has noted that for modifier-head compounds left stress may arise as a consequence of lexicalisation. Following this line of reasoning, one could argue that overprediction of right stress as seen in table 2 only appears because lexicalisation has not been taken into account. Two remarks are in order here. Plag et al. (2006a) tested the lexicalisation hypothesis, using token frequencies and orthography as two different indicators of lexicalisation. The analyses of both indicators converged on showing that indeed there is a lexicalisation effect. However, the size of the effect is very small, and, more crucially, the effect is not restricted to the categories that are predicted to be right-stressed in Giegerich's (2004) account. The problem of lexicalisation is also highly interesting from an exemplar-based perspective. Given that under such an approach all items are stored, both the would-be regular left-stressed items and the would-be irregularly right-stressed items would be available in memory and could thus serve as exemplars for analogical processes. Hence, we would expect lexicalisation effects in both directions, and not only in the direction of left stress, as Giegerich would have it.

Yet another obvious question that emerges from table 2 is whether predictive accuracies of the approaches based on argument structure and semantics could be improved if they joined forces. As is clear from the table, we cannot expect much improvement under a rule-based paradigm. Both approaches underpredict left stress, and most of those compounds for which we do
predict left stress based on argument structure are already included in the set of those compounds for which we predict left stress based on semantics.

3.2 The TiMBL and AM experiments - parameter settings

In TiMBL, highest classification accuracies were reached if similarity was computed using a simple overlap metric for the left and right member and the Jeffrey Divergence metric for all other features. Using a distance-weighted similarity metric for left and right members proved disadvantageous for the classification task, presumably because these features are represented in our corpora only as one single orthographic form. Potentially informative phonological characteristics like the number of syllables, syllable structure, rhythmic patterns were not represented. The system was thus not given suitable information that would allow it to establish similarities between different values for left and right constituents.

In experiments in which the algorithm was presented with a large number of features, classification was most successful if the distance space over which nearest neighbours were defined was set to \(k = 5\). In experiments in which fewer features were used, \(k\) was adjusted so as to make sure that the nearest neighbour set never included the whole training corpus. The voting procedure that produced best results was Inverse Distance voting. Thus, neighbours that were closer in similarity to a given test item had a greater say in classification than more distant exemplars.

In our AM experiments we had the algorithm compute analogical sets using pointers, not occurrences (cf. Parkinson, 2002 for a general outline of the options provided by AM). All items were included in the classification process. Furthermore, it is problematic in AM to use only few and abstract features to test the corpus on itself. The reason is that no leave-one-out procedure is implemented. The experimenter is only given the option to exclude a data item from consideration during classification if its context (i.e. the features) is identical to the test item. However, in a data set in which we give the system only very few features, we expect the context of our test items to be represented in the dataset. This will give us the opportunity to test, for example, whether all argument-head compounds ending in –er are successful if used as predictors in the model. Therefore, we allowed AM to include every context in the evaluation, even if this context is given in the dataset. However, we expect predictive accuracies to be higher in this case, and, thus, the result not to be directly comparable to results in the other series or to the predictive accuracy reached by TiMBL.

3.3 Series 1 – only one set of abstract features as predictor

Table 3 provides an overview of classification accuracies for the first series of experiments. We see that, in spite of the differences between the parameter settings, TiMBL and AM produce very similar classification results on the three codings. Like the rule-based accounts described in section 3.1, also exemplar-based models, if trained on argument structure and semantic categories, produce accuracies of classification hardly above chance level. Unlike rule-based accounts, however, TiMBL and AM are much better at predicting left stress than they are at predicting right stress.

<table>
<thead>
<tr>
<th>Information source</th>
<th>Accuracy overall</th>
<th>For left stress</th>
<th>For right stress</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TiMBL</strong> (with (k) adjusted so that the k-NN set can never be the whole dataset):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>argument structure</td>
<td>52.35%</td>
<td>88.27%</td>
<td>13.19%</td>
</tr>
<tr>
<td>semantic categories</td>
<td>52.22%</td>
<td>92.46%</td>
<td>12.64%</td>
</tr>
<tr>
<td>semantic relations</td>
<td>56.37%</td>
<td>56.42%</td>
<td>56.32%</td>
</tr>
<tr>
<td><strong>AM</strong> (every test item is also a member of the data set):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>argument structure</td>
<td>52.49%</td>
<td>93.3%</td>
<td>12.36%</td>
</tr>
<tr>
<td>semantic categories</td>
<td>52.77%</td>
<td>92.74%</td>
<td>13.46%</td>
</tr>
<tr>
<td>semantic relations</td>
<td>65.37%</td>
<td>65.36%</td>
<td>65.39%</td>
</tr>
</tbody>
</table>

Table 3. Classification accuracies for BURSC, series 1.
Finally, it is interesting to note that semantic relations feature differently from the other sets of abstract features. In both TiMBL and AM simulations they are the best information source. The difference between predictive accuracy of argument structure and semantic categories on the one hand and semantic relations on the other hand is statistically significant for AM (argument structure vs. semantic relations: Yate's $\chi^2 = 24.219$, $p = 0.0000$; semantic categories vs. semantic relations: Yate's $\chi^2 = 23.201$, $p = 0.0000$), but not for TiMBL.

### 3.4 Series 2 – fine-tuning the pool of abstract features

In this series we test whether we can reach better predictive accuracies if we select among the abstract features those that are most informative to the classification task and ignore the less informative ones. In order to determine which features are most informative, we pursue two different strategies. In their analysis of the BURSC data, Plag et al. (2006a) have found that only some of the abstract features produced statistically significant effects, whereas others did not. Thus, we selected these features for our experiments. They are

- argument status: only if the second constituent ends in –er
- semantic categories: 'N2 is a geographical term', 'N1 is a proper noun'
- semantic relations: 'N2 IS LOCATED AT N1', 'N2 DURING N1', 'N1 IS N2'

In a second experiment, we selected those features that TiMBL, if given all features, finds most informative for the classification task in its training phase. Unlike in AM, in TiMBL the classification of test items is preceded by a training phase, in which the model organises the features that it is trained on in terms of how informative they are for the classification task. Two different informativity measures are computed: Gain Ratio and Information Gain (for a detailed description cf. the TiMBL manual, Daelemans et al. 2004). Since Information Gain feature weighting tended to produce better classification accuracies, we used this measure as a reference. TiMBL was then trained only on the ten most informative features. They are given in table 4, together with the relevant Information Gain values.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example</th>
<th>InfoGain value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1 IS LIKE N2</td>
<td>crime wave</td>
<td>0.1021</td>
</tr>
<tr>
<td>N2 MAKES N1</td>
<td>computer company</td>
<td>0.0368</td>
</tr>
<tr>
<td>the compound is a proper noun</td>
<td>Harvard University</td>
<td>0.0317</td>
</tr>
<tr>
<td>N2 is a thoroughfare</td>
<td>state road</td>
<td>0.0206</td>
</tr>
<tr>
<td>N1 LOCATED at/in N2</td>
<td>minority area</td>
<td>0.0191</td>
</tr>
<tr>
<td>N1 is a time</td>
<td>day care</td>
<td>0.0164</td>
</tr>
<tr>
<td>N1 is a proper noun</td>
<td>Mapplethorpe controversy</td>
<td>0.0123</td>
</tr>
<tr>
<td>N1 HAS N2</td>
<td>state inspector</td>
<td>0.0077</td>
</tr>
<tr>
<td>N2 DURING N1</td>
<td>lifetime</td>
<td>0.0051</td>
</tr>
<tr>
<td>N2 LOCATED at/in N1</td>
<td>neighborhood school</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

Table 4. The 10 highest IG values in BURSC.

AM was given the same two sets of abstract features. Again, the parameters were set in such a way that all contexts were used, even if the context in the test item occurs in the dataset. Thus, a direct comparison between accuracies of classification in TiMBL and AM is not possible. The results are given in table 5.

As in series 1, the two models are very similar in their performance: Both sets of features lead to very similar predictive accuracies. In both sets left stress is much better predicted than right stress. The two models differ, however, in terms of how their predictions differ from those in series 1. TiMBL does not show any significant improvement compared to the results in series 1. AM, by contrast, is considerably better in series 2 than it was in two of the experiments in series 1 (those based on argument structure and on the semantic categories). The difference in predictive accuracy between the worse of the two tests from series 2 and the semantic categories test from series 1 is just below the level of statistical significance (Yate's $\chi^2 = 3.631$, $p = 0.0567$); for the argument structure test from series 1, the same difference is significant (Yate's $\chi^2 = 4.044$, $p = 0.0443$). However, even in AM predictive accuracies in series 2 are still not better than it was in series 1 when given only the semantic relations as features in series 1.
3.5 Series 3 – the non-abstract features as predictors

We now give AM and TiMBL only the left and the right constituents of the compounds as information source. Recall that the subset of BURSC that we are using here is set up in such a way that every test item will have a constituent family for both its left and its right constituent in the test set.

In TiMBL we carried out three experiments, testing the two constituents in isolation and in combination. In the former case, k was set to 1, in the latter case, it was set to 2, to ensure that the algorithm never used the whole training set as nearest neighbour set. Given that there is no leave-one-out procedure in AM, we had AM test only the combination of the two constituents. Since this combination is unique in the dataset, we excluded identical contexts from consideration. Thus, no item was classified on the basis of an identical context. Classification accuracies may now be directly compared between TiMBL and AM. Table 6 summarises the results.

For TiMBL, overall accuracy of prediction is optimal if a combination of left and right constituents are used as information source. In this combination, TiMBL performs better than in any of the tests in which it was trained on abstract features. Statistically, that difference in performance is significant for all combinations of abstract features tested in series 1 and 2 except for the test employing the semantic relations as predictors (for the comparison of the constituents experiment and the best combination of abstract features in series 2: Yate's \( \chi^2 = 4.988, p = 0.0255 \)). Recall that for AM, a direct comparison between series 1, 2, and 3 is not possible because in series 1 and 2 the analogical set for each test item included the item itself. Nevertheless, none of these tests is significantly better than the test that used the two constituents as information source, in spite of the fact that the tests with the abstract features had the 'advantage' that the test item was included in the dataset (Yate's \( \chi^2 = 0.149, p = 0.6996 \), comparison of the constituents experiment and the best experiment involving abstract features).

We also note that, if given left and right conf-
stituents as predictors, both TiMBL and AM are more successful in predicting right stresses than they are at predicting left stresses. This was different in most of the experiments in series 1 and 2, where for the two models' predictions of right stress were generally much worse than predictions of left stress. An exception is again the set of experiments based on semantic relations in series 1, where predictive accuracies for right and left stresses were quite balanced. For TiMBL, the difference between accuracies for right stress predictions in the semantic relations experiment and in the constituents experiment (left and right constituent) is just below the level of significance (Yate's $\chi^2 = 3.574$, $p = 0.0586$).

A comparison of TiMBL and AM in terms of general predictive accuracy in the constituents test in series 3 shows that the AM experiment yields slightly higher accuracies both in terms of overall accuracy as well as in terms of accuracies for right and left stress. However, the difference is not significant (Yate's $\chi^2 = 2.477$, $p = 0.1155$).

### 3.6 Intermediate summary

The experiments described in this section yielded a variety of important insights with respect to the nature of compound stress in English. First of all, although general predictive accuracies are not too impressive, our best TiMBL and AM simulations reached higher predictive accuracies than any of the categorical models proposed in the previous literature. The simulations employing constituent family (series 3) are significantly better than the best of the rule-based approaches introduced in section 3.1 (TiMBL: Yate's $\chi^2 = 6.542$, $p = 0.0105$; AM: Yate's $\chi^2 = 20.269$, $p = 0.0000$). Nevertheless, we should note that predictive accuracies are far from satisfactory. In this context it is interesting to note that Kunter (2007) has shown for the BURSC data that compound stress involves additional sources of variability, such as token variability within individual types, as well as type variability between different speakers. Whereas this type of variability is a challenge to all kinds of traditional, rule-based models, it is expected under an exemplar-based approach, even if it was not tested in the experimental series presented in this section.

The second thing that we can learn from the results in the previous sections is that compound stress may be computed in an exemplar-based model without assuming that abstract features are involved in the computation. A model that takes into account only left and right constituents is empirically just as adequate as a model that takes into account semantic relations, and significantly better than any other set or combination of abstract features. With respect to an evaluation of the significance of semantic relations, more research is called for. What is interesting, however, is that even this representational abstraction is not necessary, given that the 'non-abstract' constituents can do the same job equally well.

Finally, we have learned that the compound stress data provide inconclusive evidence as to which of the two algorithms, TiMBL or AM, is empirically more adequate. AM consistently performs better than TiMBL, but this difference never reaches statistical significance. Interestingly, the difference is most pronounced if the models are trained on the most relevant, non-abstract features in series 3, whereas they perform more alike in the tests involving the 'less important' sets of abstract features in series 1. This is true in spite of the fact that in series 1 AM has the advantage that every test item is also a member of the data set.

In the next session we report on a parallel set of experiments that were conducted using the CELEX lexical database. We will see that, although the database is very different, the results point into the same direction that we have seen in the BURSC data.

### 4 Modelling Compound Stress in CELEX

#### 4.1 The data, or: yet another demise of a rule-based approach

As in the BURSC experiments, we used only those compounds from the corpus that have a constituent family. The coding method was the same as for the BURSC data, except for the stress classification, which is simply a given in CELEX. As for the BURSC data, also for the CELEX compound data there exists an in-depth empirical study (Plag et al., 2006b), to which the interested reader is referred for the statistical details concerning the distributional characteristics of determinants of compound stress in the full corpus. The overall distribution of stresses is given in table 7 ($N = 2643$).
left stress | right stress
---|---
2487 | 156
94.10% | 5.90%

Table 7. Distribution of stresses in CELEX.

A major difference between the compound data from BURSC and CELEX lies in the proportion of right stresses. Whereas left and right stresses are almost equally distributed in BURSC, we find only very few right-stressed compounds in CELEX (cf. Plag et al., 2006b for discussion). As in their BURSC study, Plag et al. (2006b) tested the predictive accuracy of previous approaches on the CELEX data. The results mirror those of the BURSC study: None of the syntactic or semantic categories proposed in the literature is successful in adequately predicting the stress distribution in the corpus. Again, this is true no matter whether we implement these categories as predictors in a rule-based model or in a probabilistic, logistic regression model. Nor may predictive accuracy be significantly enhanced if features from the syntactic and the semantic approaches are combined.

For reasons of space, we will focus here only on the rule-based approaches for illustration. Table 8 summarises the predictive accuracies that Plag et al. find for the pertinent approaches, if implemented in a rule-based model.

Note that, due to methodological considerations, the figures are not based on exactly identical datasets. However, all datasets have been shown to pattern alike with respect to the pertinent features. So does the subset to be employed in the present study. As in section 3, we distinguish between overall accuracy on the one hand and accuracy for left and right stresses in the corpus on the other hand.

For argument structure we see the same effect that we saw in BURSC: The model overpredicts right stress; overall predictive accuracy does not reach chance level. For semantic relations and categories, we see an interesting difference between the predictions for BURSC and CELEX. In CELEX the approach is better at predicting left stress than it is at predicting right stress. The opposite was true for the BURSC data. The overall predictive accuracy based on semantic relations and categories is better than for argument structure.

Given the very low proportion of right stresses in the data, the CELEX data provide an extremely difficult information source for an exemplar-based model to learn stress assignment in compounds. In particular, it will be very difficult to learn the distribution of right stress on the basis of the abstract features. That this is indeed the case will be shown in our TiMBL and AM experiments in the next section.

4.2 The TiMBL and AM experiments – parameter settings

As in the BURSC simulations, TiMBL achieved best results with the Jeffrey Divergence metric as a distance metric and Inverse Distance voting among nearest neighbours. Unlike the BURSC simulations, setting the distance space to \( k = 3 \) produced better results than a higher \( k \) value. Gain Ratio was used to weight features. Again we adjusted \( k \) in experiments in which fewer features were used so that the nearest-neighbour set never comprised the whole dataset.

In AM we used the same parameters as in the BURSC settings. As in the BURSC experiments, identical contexts could be excluded from the dataset only in experiments with the two constituents as information source (series 3).

4.3 Series 1 – only abstract features

As in the BURSC simulations, we investigated whether each set of abstract features would serve as an adequate information source for TiMBL and AM. Again we distinguish between overall accuracy, accuracy for compounds that are classified as left-stressed in the corpus, and those that are classified as right-stressed. Due to the differences in parameter settings discussed in section 3.2, accuracies in TiMBL and AM cannot be compared directly.

<table>
<thead>
<tr>
<th></th>
<th>overall</th>
<th>for left stress</th>
<th>for right stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>a rule system based on...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>argument structure:</td>
<td>49.0%</td>
<td>46.9%</td>
<td>72.4%</td>
</tr>
<tr>
<td>semantic relations and categories:</td>
<td>78.7%</td>
<td>85.0%</td>
<td>30.0%</td>
</tr>
</tbody>
</table>

Table 8. Predictive accuracies of categorical, rule-based models for the CELEX data. (fro
Accuracy of classification is in general much higher than with the BURSC corpus. It is also much higher than that of the rule-based approaches sketched in section 4.1. (for the difference in predictive accuracy between the best rule and the weakest experiment in series 1: \( \chi^2 = 222.311, p = 0.0000 \)). However, it is clear that this effect is largely due to the predominance of left stresses in the training data. Both TiMBL and AM are led to predict left stress for the overwhelming majority of the data, and, because the test set is identical to the training set, high levels of accuracy may be reached just by predicting left stress. By contrast, prediction of right stresses is very weak. What this test series shows very clearly, thus, is that, even if fed the same features, an exemplar-based model may differ considerably in its predictions from a rule-based account.

It is, furthermore, interesting to note that, in spite of all problems, the tendencies that show up in the CELEX experiments are very similar to those we have seen in the BURSC experiments. In both the CELEX and the BURSC experiments, the semantic relations simulations stand out, where both TiMBL and AM manage to predict at least some right stress correctly.

4.4 Series 2 – fine-tuning the pool of abstract features

As in the BURSC experiments, we illustrate the effect that fine-tuning of the set of features given as an information source has on predictive accuracies, with two experiments. In one experiment we chose those features that proved to be significant predictors in Plag et al.’s (2006b) probabilistic model of the CELEX data. These are

- argument structure, but only for compounds ending in –er
- semantic categories: the compound is a proper noun
- semantic relations: N2 has N1, N1 has N2, N2 is made of N1, N1 is like N2, N2 for N1, N2 is located at N1, N2 is named after N1

In the second experiment we used the 10 abstract features with the highest levels of informativity for TiMBL. Since in the CELEX simulations Gain Ratio was the most successful feature weighting measure, we used the Gain Ratio values to select our features. They are given in table 10.

<table>
<thead>
<tr>
<th>feature</th>
<th>example</th>
<th>Gain Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2 IS MADE OF N1</td>
<td>stone wall</td>
<td>0.0286</td>
</tr>
<tr>
<td>N1 is a proper noun</td>
<td>India paper</td>
<td>0.0270</td>
</tr>
<tr>
<td>the compound is a proper noun</td>
<td>labour day</td>
<td>0.0220</td>
</tr>
<tr>
<td>N1 LOCATED at N2</td>
<td>telephone booth</td>
<td>0.0183</td>
</tr>
<tr>
<td>N2 FOR N1</td>
<td>writing paper</td>
<td>0.0169</td>
</tr>
<tr>
<td>N2 MAKES N1</td>
<td>silk worm</td>
<td>0.0130</td>
</tr>
<tr>
<td>N2 DURING N1</td>
<td>spring tide</td>
<td>0.0128</td>
</tr>
<tr>
<td>argument-head structure</td>
<td>fire fighter</td>
<td>0.0062</td>
</tr>
<tr>
<td>N2 is a thoroughfare</td>
<td>service road</td>
<td>0.0061</td>
</tr>
<tr>
<td>N2 NAMED AFTER N1</td>
<td>guinea fowl</td>
<td>0.0039</td>
</tr>
</tbody>
</table>

Table 10. The 10 highest GR values in CELEX.

Table 11 summarises the results of our experiments. As in the BURSC experiments, we do not see a considerable improvement of accuracies if we combine abstract features from

<table>
<thead>
<tr>
<th>information source</th>
<th>overall</th>
<th>for left stress</th>
<th>for right stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>argument structure</td>
<td>94.10%</td>
<td>99.96%</td>
<td>0.00%</td>
</tr>
<tr>
<td>semantic categories</td>
<td>94.06%</td>
<td>99.46%</td>
<td>1.28%</td>
</tr>
<tr>
<td>semantic relations</td>
<td>93.95%</td>
<td>99.46%</td>
<td>1.28%</td>
</tr>
<tr>
<td>AM (every test item is also a member of the data set):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>argument structure</td>
<td>94.10%</td>
<td>100.00%</td>
<td>7.05%</td>
</tr>
<tr>
<td>semantic categories</td>
<td>94.14%</td>
<td>99.36%</td>
<td>0.64%</td>
</tr>
<tr>
<td>semantic relations</td>
<td>94.51%</td>
<td>100.00%</td>
<td>7.05%</td>
</tr>
</tbody>
</table>

Table 9. Classification accuracies for CELEX, series 1.
different sets, even if we choose the most informative of them. We now turn to series 3, where we use the non-abstract features, left and right constituent, as information source to feed TiMBL and AM.

### 4.5 Series 3 – the non-abstract features

In parallel to section 4.5, we feed TiMBL with the left and right constituents both in isolation and in combination. Due to its limitations in parameter settings, AM is given only the combination of features. Since this combination is unique for every compound in the dataset, we could exclude identical contexts from consideration during classification.

In parallel to section 4.5, we feed TiMBL with the left and right constituents both in isolation and in combination. Due to its limitations in parameter settings, AM is given only the combination of features. Since this combination is unique for every compound in the dataset, we could exclude identical contexts from consideration during classification. The results are given in table 12.

Unlike the BURSC experiments, the constituents experiments for CELEX do not yield significantly higher predictive accuracies than those from series 1 and 2 (cf., e.g., the difference between the best experiment from series 3 and the worst experiment from series 1 and 2: Yate’s $\chi^2 = 1.893$, $p = 0.1688$). Nevertheless, the constituents experiments differ substantially from the previous series: Thus, they are the only experiments in which both TiMBL and AM manage to predict a considerable number of right stresses. The differences between the number of correct predictions of right stress between the constituents experiment and the experiment with the most predictions of right stresses on the basis of abstract features is highly significant for both TiMBL (Yate’s $\chi^2 = 17.394$, $p = 0.0000$) and AM (Yate’s $\chi^2 = 9.932$, $p = 0.0016$). The latter is true, in spite of the fact that the experiment involving 'abstract' features had the advantage that the test item itself was included in the dataset and thus enhanced predictive accuracy.

A comparable level of predictive accuracy for

<table>
<thead>
<tr>
<th>information source</th>
<th>overall</th>
<th>for left stress</th>
<th>for right stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the features found most relevant in Plag et al. (2006b)</td>
<td>94.17%</td>
<td>100.00%</td>
<td>1.28%</td>
</tr>
<tr>
<td>the 10 most informative features (TiMBL)</td>
<td>93.98%</td>
<td>99.88%</td>
<td>0.00%</td>
</tr>
<tr>
<td>AM (every test item is also a member of the data set):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the features found to be most relevant in Plag et al. (2006b)</td>
<td>94.17%</td>
<td>99.96%</td>
<td>1.93%</td>
</tr>
<tr>
<td>the 10 most informative features (TiMBL)</td>
<td>94.29%</td>
<td>100.00%</td>
<td>3.21%</td>
</tr>
</tbody>
</table>

Table 11. Classification accuracies for CELEX, series 2.

In parallel to section 4.5, we feed TiMBL with the left and right constituents both in isolation and in combination. Due to its limitations in parameter settings, AM is given only the combination of features. Since this combination is unique for every compound in the dataset, we could exclude identical contexts from consideration during classification.

In parallel to section 4.5, we feed TiMBL with the left and right constituents both in isolation and in combination. Due to its limitations in parameter settings, AM is given only the combination of features. Since this combination is unique for every compound in the dataset, we could exclude identical contexts from consideration during classification. The results are given in table 12.

Unlike the BURSC experiments, the constituents experiments for CELEX do not yield significantly higher predictive accuracies than those from series 1 and 2 (cf., e.g., the difference between the best experiment from series 3 and the worst experiment from series 1 and 2: Yate’s $\chi^2 = 1.893$, $p = 0.1688$). Nevertheless, the constituents experiments differ substantially from the previous series: Thus, they are the only experiments in which both TiMBL and AM manage to predict a considerable number of right stresses. The differences between the number of correct predictions of right stress between the constituents experiment and the experiment with the most predictions of right stresses on the basis of abstract features is highly significant for both TiMBL (Yate’s $\chi^2 = 17.394$, $p = 0.0000$) and AM (Yate’s $\chi^2 = 9.932$, $p = 0.0016$). The latter is true, in spite of the fact that the experiment involving 'abstract' features had the advantage that the test item itself was included in the dataset and thus enhanced predictive accuracy.

A comparable level of predictive accuracy for

<table>
<thead>
<tr>
<th>information source</th>
<th>overall</th>
<th>for left stress</th>
<th>for right stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>TiMBL (with k adjusted so that the k-NN set can never be the whole dataset):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left constituent</td>
<td>94.32%</td>
<td>98.87%</td>
<td>21.79%</td>
</tr>
<tr>
<td>right constituent</td>
<td>93.27%</td>
<td>97.91%</td>
<td>19.23%</td>
</tr>
<tr>
<td>left and right constituent</td>
<td>94.32%</td>
<td>99.32%</td>
<td>19.23%</td>
</tr>
<tr>
<td>AM (the test item is never a member of the data set):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>left and right constituent</td>
<td>94.85%</td>
<td>99.56%</td>
<td>19.87%</td>
</tr>
</tbody>
</table>

Table 12. Classification accuracies for CELEX, series 3.
right stress is only found in the rule-based approach based on semantic relations and categories that was described in section 4.1. (30.0% accuracy for right stress). The categorical model, however, reaches its high level of accuracy by overpredicting right stress, which results in a relatively poor predictive accuracy for left stress (85.0%) and, as a consequence, an overall predictive accuracy that is significantly lower than the simulations in series 3 (cf., e.g., Yate's χ² = 237.595, p = 0.0000, comparison between the TiMBL simulation with two constituents and the rule-based model based on semantic categories and relations).

Finally, as in all other experiments, we do not find a significant difference between the performances of TiMBL and AM in series 3.

4.6 Intermediate summary

In terms of overall predictive accuracy, our exemplar-based models significantly outperform the categorical, rule-based approaches that are proposed in the literature. In this respect, the CELEX experiments very much resemble the BURSC experiments presented in section 3. However, the CELEX experiments differ from the BURSC experiments in that in most simulations, both TiMBL and AM grossly overpredict left stress. Strikingly, the one series of experiments in which both models predict at least a substantial amount of right stresses is the series that uses only the two 'non-abstract' features, the constituents, as an information source. In this respect, the experiments in series 3 are more successful than those employing abstract features (series 1 and 2), although, due to the large proportion of left stresses in the data, these differences do not result in statistically significant differences in terms of overall predictive accuracy.

With respect to the question of whether differences between nearest neighbour selection in TiMBL and analogical set composition in AM in terms show in experiments on compound stress, we are presented with a similar picture as in the BURSC experiments: AM’s levels of overall accuracy are consistently above those reached by TiMBL, but this difference is well below statistical significance.

5 Conclusion and Outlook

The corpus-based empirical study of English compound stress has brought to light a wealth of evidence in favour of an exemplar-based model of compound stress. In terms of predictive accuracy, we have shown that, if trained solely on the right and left constituents of each compound, computational models like TiMBL or AM are more successful in predicting the locus of stress than any categorical rule that has been proposed in the literature, including the Compound Stress Rule. More specifically, the predictive accuracy of – the would-be ill-behaved - right-hand stress improves if the model is trained on the left and right constituents only. Thus, our findings suggest that the level of abstraction needed to compute stress in compounds over stored exemplars does not require abstract syntactic and semantic features. A representation of left and right constituents suffices. This is in line with recent findings concerning the role of constituent families in compound semantics and compound morphology (e.g. Gagné, 2001, Krött et al., 2001, 2002). It is also paralleled by other recent studies, which argue that word stress assignment is influenced by stress of phonologically similar items in the lexicon (cf., e.g., Daelemans et al. 1994, Eddington 2002 for computational models, Guion et al. 2003 for experimental evidence on English), but contrary to much work in metrical phonology, where it is generally assumed that stress is part of the lexical representation of individual items only in cases of 'exceptional' stress assignment.

Comparing the performance of TiMBL and AM, we see that there are no significant differences in predictive accuracies. However, we also note that AM is consistently a little better than TiMBL in almost all experiments. Interestingly, at least for the BURSC simulations that difference is most pronounced in those experiments in which the model is given features as information source that it finds useful. Conversely, the difference almost disappears if the model is given less useful features (e.g. argument structure, semantic categories), although in these experiments even an exact copy of the test item is included in the dataset. If these observations can be substantiated in more detailed testing, the observed differences between AM and TiMBL may provide additional support for our claim that the the non-abstract constituents are better predictors of compound stress than the pertinent abstract features. Studies comparing TiMBL and AM (e.g. Eddington, 2002, Krött et al., 2002) have mentioned that TiMBL is less influenced by noise or irrelevant features in the data than AM. If abstract features are irrelevant, then we expect AM to have prob-
lems if fed with these features, and this is indeed what we find.

Finally, we need to emphasise that research in exemplar-based modelling of compound stress cannot stop here. In this paper we have presented only the first, necessary step. Although TiMBL and AM outperform the categorical and statistical models discussed in Plag et al. (2006a, 2006b), they still produce a considerable amount of classification errors. At this point it is important to note that our TiMBL and AM simulations have neglected two aspects which many exemplar-based approaches to linguistic generalisation consider a vital ingredient of exemplar-based modelling, but whose incorporation is far beyond the scope of this paper.

- the continuous classification of test items
- the influence of frequency factors

In our experiments we have set TiMBL and AM to the task of categorically classifying each test item as either left-stressed or right-stressed. This, however, is an idealisation of the real facts. Using case studies from the BURSC corpus, Kunter (2007) shows that, contrary to prevalent tacit assumptions in most of the pertinent literature, there is inter- as well as intra-speaker variability of stress assignment in compounds. Crucially, compounds differ in terms of the extent to which they exhibit such variability. These facts support an exemplar-based approach to compound stress where different individual realisations of a single compound are assumed to be stored in memory. The question that arises, then, is if we can enhance predictive accuracy by having our test items classified continuously. Whereas both TiMBL and AM provide parameters which allow us to get continuous classification as outputs, however, neither the BURSC nor the CELEX data are suitable to assess token variability in compound stress beyond the case studies analysed in Kunter (2007).

Secondly, we have exclusively relied on types in our experiments. However, in line with much of the literature in exemplar-based modelling, we may expect different types of frequency to play a role in determining the strength of individual exemplars. Among the candidates to be tested are the token frequencies of each compound as well as the family size of the constituents of our compounds. Starting from where this paper ends, it is a task for future research to test whether the integration of both variability and frequency factors into our exemplar-based model of compound stress considerably improves classification.

References


Kunter, G. (2007). Within-Speaker and between-Speaker Variation in Compound Stress Assign-


Accounting for Phonetic and Syntactic Phenomena in a Multi-Level Competitive Interaction Model

Michael Walsh  Hinrich Schütze  Bernd Möbius  Travis Wade
Institute for Natural Language Processing
University of Stuttgart
Germany
firstname.lastname@ims.uni-stuttgart.de

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 over-arching 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.

1 Introduction

Exemplar-theoretic models are among the most successful in explaining human categorization (Nosofsky, 1986; Nosofsky and Zaki, 2002). There is also an increasing body of work applying exemplar-theoretic models to phonetic phenomena (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. Exemplar theorists posit that language comprehension and production are achieved via operations on these stored traces. Thus, when a new exemplar is encountered it is classified on the basis of its similarity to previously stored exemplars.

The appeal of exemplar models 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 et al., 2003) and frequency effects in phonetics (Jurafsky et al., 2001) and syntax (Bybee, 2006).

Our main contribution in this paper is that we present a unified exemplar 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 mod-
els (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.

It is also worth noting that our competition model is in keeping with dual-route theories of speech encoding where linguistic events frequently encountered by speakers are processed in a different manner to those which occur infrequently (Levelt and Wheeldon, 1994). In essence, a direct-route production corresponds to the production of a stored plan (e.g. a stored syllable plan in a mental syllabary) and an indirect route corresponds to a production-via-assembly mechanism. The intuition here is that direct-route encoding represents intelligent storage of high frequency forms and is likely to facilitate efficient error free production. Empirical evidence for a dual-route hypothesis comes from differences in duration between high and low frequency speech forms, with high frequency forms having shorter durations than lower frequency cognates (Varley and Whiteside, 1998).

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. Up to now little work has been carried out on formal exemplar-based syntactic models as it was unclear how infinite productivity could be achieved (but see Section 6).

One of the important theoretical questions in exemplar theory concerns the status of prototypes, where a prototype represents an extant exemplar, or an exemplar representation, which holds a special status of being indicative of the category. It has often been argued that a purely exemplar-theoretic account, i.e. one where prototypes are not employed, fails to explain a number of observations in human categorization (e.g., during early learning of a category, (Smith et al., 1997)). The model proposed here is strictly exemplar-theoretic without any prototype component. We will show that for the three phenomena that we are concerned with, prototypes are not needed.

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 syllable 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 and related work and Section 7 offers some 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.** This component generates (possibly underspecified) unit exemplars that serve as stimuli for the model. It is 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.
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 (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 syllables. The first simulation tests whether we can reproduce these experimental findings.

#### Assumptions.

The model is based on two assumptions. First, we assume that syllables have a wide range of different frequencies, with frequent syllables being several orders of magnitude more frequent than infrequent syllables. For the experiment, we choose a factor of 100, which we found to be not uncommon in previous work (Müller et al., 2000; Müller, 2002).

Secondly, previous research has shown that a significant part of the variation in syllable durations can be explained by variation in the duration of their constituent segments (van Santen and Shih, 2000). Our interpretation of this finding is that, unless a syllable is produced as a whole (in the way suggested by Schweitzer and Möbius’s model for frequent syllables), a syllable’s duration is closely approximated by the sum of the durations of its constituent segments.

#### Stimuli.

Stimuli were syllables of the form CVC

---

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.
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).

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:

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

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

where $x$ and $y$ are durations and $\alpha = 0.05$. $\alpha$ 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 7), 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, then with probability 0.99, it is discarded 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 $\theta_1$ (i.e., there are fewer than $\theta_1$ 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 mem-
Figure 2: Experimental results for variation of syllable duration. Infrequent syllables (dashed line) have lower variability in duration than frequent syllables (solid line).

numbers of its cloud. Finally, random noise proportional to the computed duration is added. The choice of the radius parameters and of $\theta_1$ will be discussed below.

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 the probing phase, syllables and their units are deleted after each probing to make sure that infrequent syllables do not change their status to frequent in this phase.

**Results.** Figure 2 is a cumulative histogram of 10 runs of the above experiment, corresponding to 1250 standard deviations. The model successfully simulates the finding of Schweitzer and Möbius (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 by 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 contexts of words 1 and 2, respectively, and $D_{KL}$ is the Kullback-Leibler divergence (and analogously for the right contexts of two words). We do not use KL divergence directly as a distance measure because it is asymmetric and undefined if there are words that do not occur in one of the two contexts (because of zero probability values).

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 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 in-
ferred 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 judgements

One of the basic tasks children master when acquiring a language is to distinguish between grammatical and ungrammatical sentences. Rote learning is no help in judging grammaticality because of the productivity of language. 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, on the level of syntax, 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.

We restrict our model to a subset of three-word sentences in the early stages of language acquisition. In particular, we do not attempt to model the acquisition of recursive phenomena (as, e.g., (Klein and Manning, 2004) do). While there have been many previous models of syntax acquisition, none has been exemplar-theoretic, to our knowledge.

It is also important to point out that the acquisition of three-word sentences is trivial for a system that has full knowledge of syntactic categories. If after a few hundred stimuli, only subject-verb-object sentences have been observed and no subject-object-verb sentences, then rote learning is sufficient to predict grammaticality for new utterances correctly.

However, the acquisition of syntactic categories goes hand in hand with the acquisition of grammaticality in child language acquisition. A complete model needs to account for the parallel acquisition of both without assuming the prior existence of either. Our model provides such an explanation and does so within an exemplar-theoretic framework.

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 successfully models the acquisition of grammaticality of three-word sentences 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). Although the learning taking place here is with respect to a small subset
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).

of English, generalising to larger left and rights contexts should not prove problematic. In addition it is important to note that, as with the previous two experiments, the same model of unit and constituent interaction is employed here.

6 Discussion

6.1 Abstractionist models

We have presented an exemplar-theoretic model that makes correct predictions for three linguistic phenomena. It is noteworthy that the model achieves this without prototypes or any explicit abstraction mechanism. At least for the three phenomena investigated here, a simple exemplar model without prototypes seems to be sufficient. Note, in particular, that Abbot-Smith and Tomasello (2006) express doubts that a pure exemplar-theoretic model can account for grammaticality judgments in early child language acquisition. With respect to exemplar models they hold the view that each comprehension of an exemplar must, minimally, result in a change in its representation (even if this is a simple recording of frequency). Furthermore, they also propose that frequent summing over mutual similarities of a particular cloud of exemplars is “highly likely” to result in a permanent modification to the representation which is “in some way equivalent to the formation of some kind of more abstract representation” (Abbot-Smith and Tomasello, 2006). In other words, the hybrid categorization model which they propose allows for exemplar learning and retention but also offers an abstraction mechanism where a more abstract schema is somehow encoded in the summed similarities. However, while the comprehension of an exemplar might strengthen the activation of an exemplar cloud as a whole, this does not necessarily entail that the exemplar representations themselves have to change. Indeed, the model presented here illustrates accurate syntactic acquisition without the need for any modification of stored exemplars nor any form of more abstract representation. That is novel stimuli are correctly categorized through direct comparison with extant exemplars. Thus, for the three disparate phenomena examined above, exemplar theory appears to provide an adequate account. While it could perhaps be argued that some form of abstraction is implicitly encoded in the summed similarities in our model, there is certainly no explicit abstraction component.

6.2 Radius-based vs. nearest-neighbor models

In both production and comprehension, exemplar-theoretic models infer the property of a stimulus from the properties of exemplars that are similar to the stimulus. A fundamental question is therefore how the set of relevant similar exemplars is to be computed. This set can be either defined as those exemplars that are within a fixed radius \(d\) (radius-based model) or as the set of the \(k\) nearest neighbors (nearest-neighbor model) where \(d\) and \(k\) are parameters.

We can regard the degree of activation (or resonance) a stimulus receives as part of the inference process. Activation is high if many similar exemplars exist. It is low if the most similar exemplars are distance or if there are only a few highly similar exemplars. Radius-based models support a simple definition of activation: the number of exemplars in the relevant exemplar cloud (that is, all exemplars that are at a distance of at most \(d\) from the stimulus).
It is more difficult to define activation in a nearest-neighbor model since, by definition, there are always $k$ nearest neighbors. One could attempt to derive a measure of activation by weighting neighbors according to similarity. However, the resulting model would not be a true nearest-neighbor model, but a hybrid that would need to specify which distances are still considered close enough to give rise to high weights.

The notion of activation in a radius-based model is crucial for all three simulations presented in this paper. In the phonetic model, even an infrequent syllable has $k$ nearest neighbors. Thus, when making the decision as to whether there is enough activation for the syllable to be produced by the unit model, it would not be clear how to distinguish frequent and infrequent syllables. In contrast, the distinction is straightforward in the radius-based model we presented. Similarly, the difference between the grammaticalization of *going to fetch* vs. non-grammaticalization of *walking to fetch* requires the same notion of activation: The former neighborhoods are “denser” because of the high frequency of “go” compared to “walk. Finally, in the case of grammaticality, even ungrammatical sentences have nearest neighbors (albeit neighbors that are far away). Again, it is not clear how grammaticality judgments could be modeled with nearest-neighbor clouds.

In our opinion, the experiments show conclusively that neighborhoods in exemplar theory must be radius-based as opposed to nearest-neighbor. Previous arguments in favor of nearest-neighbor clouds were based on difficulties found in implementing fixed-radius models (Pierrehumbert, 2001) and not on any fundamental reasons.

### 6.3 Multi-level models

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. We believe that we have provided the first exemplar-theoretic model that explicitly models constituency, either at the level of phonetics or syntax. Furthermore, our research represents a first step towards placing syntactic exemplar theory on a more formal footing with explicit statements of the assumptions of the model and the ability to test them against data.

Up to now, the majority of exemplar-theoretic work on syntax has been informal (e.g., (Abbot-Smith and Tomasello, 2006; Bybee, 2006)). However, Bod (2006) has recently argued that data-oriented parsing (DOP) is an exemplar model. There are significant differences, however, between DOP and standard exemplar theory. In particular, the exemplar cloud in DOP is a superset of the set of all sentences that have one or more words in common with the stimulus. No notion of similarity between the stimulus and one of the members of its exemplar cloud is defined. Hence, DOP appears to lack features which are central to most exemplar models. To its credit, however, no formal exemplar model offers such a full exemplar-based account of grammatical productivity as DOP provides. We anticipate that when current informal models (e.g. those of Bybee (2006) and Abbot-Smith (2006)) are formalized, much progress will be made because implicit assumptions will become explicit, and predictions testable against real data.

### 7 Future Work

One aspect of the work we have presented here which could benefit from further examination is the manual selection of the parameters $d$ (the radii of the exemplar neighborhoods) and the thresholds $\theta$ (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 $\theta$). 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 than those 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.

8 Acknowledgments

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

References


Modelling the steps of early syntax acquisition

Jacqueline van Kampen
UiL OTS
Utrecht University
Utrecht, 3512 BL
jacqueline.vankampen@let.uu.nl

Remko Scha
ILLC
University of Amsterdam
Amsterdam, 1018 TV
scha@uva.nl

Abstract
This paper reports on empirical research concerning the early stages of syntactic development in first language acquisition, and investigates the prospects of modelling the acquisition process in an exemplar-based linguistic framework (Data-Oriented Parsing). We present a detailed analysis of the early stages of verbal placements in Dutch. Our analysis shows that the child invents a sequence of grammars. She starts with rigorous, but systematic simplifications. The child approaches the adult grammar in a sequence of steps. We propose that each acquisition step can be characterized by (a) a new intermediate grammar and (b) a corresponding input reduction strategy that selects the new data for the next acquisition step. We discuss in some detail the how a child acquires the complexities of Dutch word order. It turns out to be extremely implausible that this process could be simulated by a model whose syntactic operations are limited to tree-substitution.

1 Introduction
This paper consists of two parts. In the first part (section 2), we discuss some complexities of Dutch word order, and some empirical findings about the stepwise process by which a Dutch child ends up introducing these complexities in the language that it is developing.

In the second part of the paper (section 3) we go into more detail about the problem of modelling this acquisition process. In particular, we investigate to what extent exemplar-based linguistic approaches such as Data-Oriented Parsing (DOP) provide suitable frameworks for this purpose.

2 Acquiring Dutch word order
Dutch clauses come in different types, each with its own word order. A Dutch three year old typically masters all the word order variants in (2)-(8). The ‘root infinitive’ type in (1), an invariant moment of the variants in (2)-(8), is as such not present in the adult input. Nevertheless, it is present in the child’s output and precedes the types in (2)-(8).

(1) ‘root infinitive’
   a. Minou muis bijten
   b. Minou mouse bite

(2) root statement
   a. Minou gaat nu in de muis bijten
      Minou will now in the mouse bite
   b. Minou bijt nu in de muis –
      Minou bites now in the mouse

(3) question
   a. gaat Minou nu in de muis bijten?
      will Minou now in the mouse bite?
   b. bijt Minou nu in de muis –?
      bites Minou now in the mouse –?

(4) imperative
   a. ga nu in de muis bijten, Minou!
      go now in the mouse bite, Minou!
In each variant, the verb *bijten* with its arguments *Minou* and *in de muis* is placed in a different grammatically marked context and the order of verb and arguments changes accordingly. The systematic similarity that binds the variants together is each time the verb *bijten* with those two arguments *Minou* (agent) en *de muis* (goal). The variant requirements of sentence-typing interact with an invariant set of lexical core elements (argument – theta frames). The theta frames bring in a lot of idiosyncratic restrictions that are characteristic for the specific language.

Child language clearly reconstructs the patterns (2)-(8) in a stepwise fashion. The present paper will trace the sequence of intermediate steps that lead from the one-word stage to the adult language.

Van Kampen (1997) described a few actual developments in child Dutch as an ordering of grammars $G_0,...,G_i,...,G_n$ offering an explanation for that order. Child language clearly employs a data-reduction procedure. Initially, the child ignores all functional categories. These functional categories are successively picked up from the input in a stepwise reconstruction of the adult input. This leads to the first “syntax”, which combines two content words in a topic+comment structure (*beer slapen* ‘bear sleep’, *beer lief* ‘bear nice’) or an [illocution-operator + comment] structure (*moet slapen* ‘must sleep’, *is lief* ‘is nice’).

In the next stage, these two-word structures get nested, as in (9) and (10).

(9) beer slapen + moet slapen →
   a. beer [moet slapen] (bear [must sleep])
   b. moet [beer slapen] (must [bear sleep])

(10) Minou bijten + muis bijten →
    Minou [muis bijten] (Minou [mouse bite])

What one sees in child language is the insertion of new functional categories (grammatical markings) within existing frames next to rearrangements of subparts. There is an order of appearance for all innovations and an acquisition model should be able to explain that order. Some functional categories are acquired before others and by being earlier they serve as an acquisition basis for later ones.

The transition in child language from two-word structures (*beer slapen*) towards constructions that merge or rearrange preexisting phrases (*moet – beer slapen*, *beer – moet slapen*) enables the acquisition of grammatical categories in basic context. Besides these constructions, that suggest a basic frame of functional categories that allows lexical substitution, there seems to be a more advanced and abstract variant that suggests meaningful changes in the order of categories, rather than a mere lexical substitution in a fixed category frame.

The transition in child language from two-word combinations and merged re-arranged constructions towards abstract movement structures, can be demonstrated by considering the acquisition of the Dutch V-second pattern, see (11).

(11)
   a. two-word ‘root infinitive’
      *beer slapen* (bear sleep)
   b. merge illocution operator <+aux>
      *gaat beer ook slapen* (goes bear also sleep)
      *beer moet ook slapen* (bear must also sleep)
   c. move finite verb
      *beer slaapt ook* – (bear sleeps also –

One might wonder why (11)c could not be acquired as a new construction not related to (11)a or (11)b. A more detailed consideration of the acquisition course of (11)c in the next subsection will
show that the relation with types (11)a and (11)b is clearly present.

2.1 An acquisition paradox: the instantaneous acquisition of subordinate word order

About the acquisition of the adult word order in Dutch main clauses and subordinate clauses, we may note the following paradoxical observations:

I. It is known from comparative grammar that the grammatical forms of the main sentence are more varied than those of the subordinate. Moreover, the latter are simpler and closer to the invariant properties. This is sometimes referred to as the Penthouse Principle (Ross 1973: “upstairs”, i.e. in the main sentence, there are more facilities than “downstairs”, i.e. in the subordinate). The child's input thus contains several different types of main clauses. Yet, none of the main clauses realizes the pattern required for the subordinate.

II. The stimulus for the root pattern is remarkably rich. Though 95% of the child's input sentences have a finite verb in first or second position, the finite verb is systematically left out in early child language (disregarding a few stereotypical items that will be verbs later on appear in the clause final position. This is the so-called “optional infinitive” stage, (Minou bijten, mais bijten). As we noted above, a second step towards the adult input is the insertion of (finite) modals, aspectuals or the copula in the first or second position. As a last step, the finite denotational verb appears in the V-second position, leaving a gap in the predicate, (Minou bijt mais –.). This step takes the child some 15 weeks and roughly a million elementary experiences. It is shortly thereafter (well before the third birthday) that the finite subordinates first appear.

III. The stimulus for the subordinate pattern is remarkably poor, no more than 2% of the input. The subordinate clause pattern has the finite verb in the predicate-final position, see (8)b. This pattern is not realized in the root clauses of the preceding acquisition stages.

And here is the paradox: Although subordinates constitute less than 2% of all input sentences (Van Kampen 1997: chapter 2), the acquisition of the subordinate pattern is instantaneous. The subordinate pattern contradicts the hard-won and massively reinforced pattern of the root sentence.

The question now is: how can a model of language acquisition account for the instantaneous acquisition of the subordinate order in spite of (i) the pattern reversal, (ii) the high input frequency of the main clause, and (iii) the almost negligible input frequency of the subordinate?

The answer to this question lies in the order of the acquisition steps. The acquisition of the verb-second pattern itself was based on a previous acquisition step. All denotational verbs in Dutch have first been acquired as <!finite> forms in the predicate-final position. That previous lexical learning step must have been solidified in the lexical frame for verbs. All Dutch verbs are associated with a lexical argument-verb frame and within that lexical frame the verb takes the final position. It now follows that the subordinate pattern is immediately supported by the lexical frame of the verb, whereas the finite verb requires a change of that frame. The effect of an acquisition stimulus therefore depends on the acquisition steps that have been taken before.

We must conceptualize language acquisition as a linear array of separate acquisition steps arranged in such a way that the later steps are possible only due to the earlier steps. The questionable point of such a reconstruction effort is that the parental input offers all grammatical markings and distinctions simultaneously. Hence, the acquisition procedure must make a selection from the material of a (still) non-analyzable input. Yet, this is possible. The acquisition procedure opens de facto with a thorough but successful reduction on the input.

2.2 Input reduction and its effects

Child language and especially early child language, is reduced with respect to the parental input. A simple common sense principle for the child must be: ignore or leave out what you cannot fit in. Before any syntax is acquired, this will result in leaving out at least all grammatical words (functional categories). The residue will consist of a few, ideally two, words that have an immediate situation-bound meaning or function. The relation between such situation-bound words may be captured pragmatically. If the parental input connects these two situation-bound words with a grammatical marker, such an element can be acquired as an optional approximation of the adult paradigm. The optional addition of such a grammatical word will
become more frequent and ultimately develop into the consistent use of a functional category.

In the next stage, the reduction procedure will reapply to the input with a less dramatic effect. Functional categories that have become identified will pass the reduction filter, and the acquisition procedure may shift its attention to a new functional category or marking that may be added. In principle, the input reduction allows the acquisition procedure to scan the input for the best identifiable grammatical words and markings. Grammatical systems and their use must be such that they allow such a stepwise “decoding procedure”. We may model this as in (12).

(12) Acquisition scheme
a. substitute <F?> for all unidentified grammatical markings
b. remove all input sentences containing more than one <F?> (Single Value Constraint, Berwick 1985).
c. identify the cognitive/pragmatic correspondence of all constructions marked by the same <F?> as the meaning/ function of <F?> → F₁ and reapply the procedure.

Applied to the verbal distributions of Dutch, this will lead to the following four acquisition steps. First step ((13)a). Children learn at first <-finite> structures. On the basis of input such as gaat Minou de muis bijten? (goes Minou the mouse bite?), the child constructs utterances such as Minou muis bijten (‘Minou mouse bite’). The <+finite> verbs are ignored.

Second step ((13)b). Shortly afterwards, auxiliary verbs like gaat (‘goes’) are added: gaat Minou muis bijten. (‘goes Minou mouse bite’)

Third step ((13)c.) Within a few months, the <+fin> denotational verb is inserted at the position of the auxiliary: Minou bijt muis (‘Minou bites mouse’).

Fourth step ((13)d). The subordinate appears with the finite verb in predicate-final position (als Minou de muis bijt (‘if Minou the mouse bites’).

(13)

a. [∅ ----- [ Vnon-finite ] ]
(M.) muis bijten

b. [Auxfinite ----- [ Vnon-finite ] ]
(M.) gaat muis bijten

c. [Vfinite ----- [ ∅ ] ]
(M) bijt muis

d. [Vfinite ----- [ Vfinite ] ]
(als M.) muis bijt

The four steps are described in (14), with the input percentages added.

(14)  
a. Root infinitives. Nearly all finite verbs are left out ((13)a; in spite of 100% counter-evidence)

b. Insertion of the illocution sentence type operator <+aux> ((13)b covers > 66% of the input)

c. Filling the illocution operator gap by movement of the finite verb ((13)c covers < 34% of the input)

d. Recognizing subordinates as non-illocutions ((13)d covers < 2% of the input)

The four steps are as a matter of fact quantitatively present in the CHILDES files as successive stages. Consider the acquisition graph of Dutch Sarah (Evers & Van Kampen 2001, 2007) in (15).

(15) The rise of finite verbs in V-second position (Dutch Sarah)
The instantaneous acquisition of the subordinate order may now be explained by a system-internal development that is data-driven. This process consists of five steps:
(i) The lexicon records the underlying verbal position as predicate final due to the initial reduction of all functional elements, including finite verbs. 
(ii) The predicate-final position of the verb enters into the lexicon and offers the exemplar-based frame for the subcategorization of predicate heads/verbs. Fortunately, it has been shown that the linear argument structure can be identified as a subcategorizing argument frame that is associated with the predicate heads (Brent 1994). We now propose that (13)b and (13)c around week 115 (at the 60% level) can still be modeled as “alternative exemplars”. As soon as the overlap develops into a substantial percentage of the adult norm at week 122 (the 80% level), each denotational verb has some 1/3 (34%) chance to appear as <+finite> and 2/3 chance to appear as <−finite>.
(iii) The set of items that allows this variation defines the category <+V>. Each <+finite> pair shares the same argument frame.
(iv) The fact that the denotational V <+finite> corresponds with an empty place in predicate-final position, is no longer something that is noticed for each item separately. It is a categorial property of the Dutch verb.
(v) The learnability of subordinates follows now as a simplified version of the earlier pattern (13)c.
If the pattern shift (13)c is categorial rather than item-by-item, one expects the points in (16).

(16)

a. absence of exceptions, verb-like items that lack a +finite> or <−finite> variant.
b. sudden rise of verbs that appear as <+finite>.
c. the instantaneous acquisition of the subordinate pattern (13)d.

An acquisition model needs something like rearrangements to represent the succession of the exemplars (13)b,c,d as a system-internal development.

This observation about the acquisition process fits an observation about the adult language. As we saw in (2)-(8) above, there is a variety of verb-argument patterns in Dutch; the choice depends on the type of clause involved. A grammatical model that does not provide a verb movement operation cannot describe the systematic relations between these patterns. It would not be able to derive a main-clause pattern from a subordinate-clause pattern (or vice versa). It would therefore be forced to learn the patterns one by one for each combination of verb and clause-type. Such an approach might ultimately be able to reproduce the distributional facts, but it would fail to draw any advantage from the exception-less regularity of these patterns: in Dutch main and subordinate order are systematically different without any lexical exception. It would especially fail to predict the instantaneous character of the actual acquisition step.

The variation between root and subordinate exemplars can be captured if they are related by derivational rules. Interestingly, the acquisition procedure itself betrays that, due to an initial reduction of the input strings, exemplars displaying the S-O-V<−finite> structure are stored before the derived exemplars displaying the S-V <+finite>−O order appear.

### 2.3 An acquisition procedure

An empirically adequate acquisition algorithm should simulate the acquisition of V-second along the lines above. It should select the data without a priori guidance and predict the actual order of acquisition steps. It would, for every new input string (i) perform the input reduction.
(ii) select a frame for a new functional category, such that:

a. the syntactic/morphological form and the semantic/pragmatic form of the category is identified
b. the match between actual input and reduced intake is improved
c. a result can be stored in the acquisitional memory base of the model

By processing a large enough corpus of child-directed utterances, the program should be able to derive all the grammatical markings, and at least the V-second and subordinate patterns. The model may support the following perspective:

a. The actual order of acquisition steps reveals a hierarchy between less variant and more variant properties of the structures of the language.
b. Clause structure is a matter of transformationally deriving the sentence typing and variant parts of the structure from the lexical invariants.
c. Functional categories that trigger movement transformations allow a simpler lexicon and a short-cut to the acquisition of the variety in grammatical patterns.

3 Language acquisition and Data-Oriented Parsing

3.1 DOP

Scha (1990, 1992) argued that human language processing crucially employs a memory for previously processed analyses and their occurrence frequencies: a new input string gets the analysis that displays the simplest analogies with the largest number of stored structures. So far, computational models based on this approach usually employ some form of Stochastic Tree Substitution Grammar (STSG) (cf. Bod 1995; Bod, Scha & Sima'an 2003).

Data-Oriented Parsing differs from conventional STSG in its choice of the set of elementary trees of the grammar. This is a very large and redundant set: the sub-trees of all utterance-analyses in an annotated corpus that is assumed to represent the "past linguistic experiences" of the language user that is being modelled.

In a language model of this sort, a person's "knowledge of language" is not summarized in a concise grammar. It is distributed across a large, redundant data set that allows competing sets of rules and exceptions. The model is "exemplar-based" in that all successful analyses are stored and may have an influence on future processing. But it has the same kind of recursive productivity as a conventional grammar, because every exemplar implies all its abstractions.

For instance, if a child-language corpus includes the exemplars

S:[topic: beertje, pred: slapen]
S:[topic: α, pred: β]
S:[topic: α, pred: β]
S:[topic: α, pred: [patiens: χ, action: bijten]]
S:[topic: α, pred: [patiens: χ, action: δ]]

the subtree-set that is used to analyze new input (the elementary trees of the STSG) also includes, for instance:

Note that these abstractions from the exemplars include the subcategorization frames of the verbs (which in a conventional grammar would be stored in a "lexicon"), as well as completely abstract structures that in a conventional grammar would be called (context-free) rules – and indeed, they function as such. The important difference with a conventional frame or a conventional CFG-rule lies in the fact that in a DOP-model the abstract structures must always compete with more completely lexicalized tree-fragments.

What may be substituted for α, β, χ, δ depends on the intersubstitutability-classes ("categories") that have been developed at this stage.

The redundancy of the DOP approach has an advantage when we model the process of language acquisition: new constructions can emerge gradually and supersede existing ones. Language acquisition itself may be viewed as an evolutionary process. Certain patterns may become dominant and reanalyze the domain of their competitors. How this idea pans out in practice is not yet known. We need both a computational model that implements it and an assessment of its compatibility with the empirical facts of language acquisition. Below we explore the prospects of such an enterprise.

3.1 Modelling the acquisition of Dutch verbal placements

The earliest stage of a person's linguistic development is sometimes viewed as particularly problematic for the data-oriented approach, since at that point there is as yet no corpus of previously analyzed input. How does language acquisition begin? This question highlights a semantic/pragmatic component of language processing which is to some extent independent of syntax. A language user always tries to match the semantic/pragmatic context of an utterance with its possible interpretations. In the early stages of language use, this component must be the dominant one. Later, the
influence of past language experience becomes increasingly prominent (cf. Scha 1990).

Consider, for instance, the stereotype topic-predicate meanings of the child's first word combinations. Constructions like [beer slapen] (‘bear sleep’) [pappa lief] (‘daddy nice’) are built up by connecting a name with a word for a situation type, borrowing saliently stressed words from the adult utterances. A computational model demonstrating this process is presented by Chang & Gurevich (2004).

In the present paper we focused on a later stage in the acquisition of Dutch. After two-year olds have built up a corpus of one-word and two-word sentences they go through an interesting syntactic development. They insert modals and copulas between the topic and the predicate, thus moving from stage (13)a to (13)b. The new structure (13)b is extended later on by the introduction of finite denotational verbs.

In the adult speech that the child encounters, the surface word order patterns (17)a,b,c occur, whereas the patterns (17)d,e do not occur.

(17)

- a. S Aux OV< finite> Minou gaat muis bijten
- b. S V< finite> O Minou bijt muis
- c. V< finite>SO nu bijt Minou muis
- d. *S Aux VO
d. *Aux S VO

The sentence types (17)a,b,c are all in the input, and all allow the type assignments for subject, object and (denotational) verb, but due to the large number of auxiliaries (>66%, see (14)) the pattern [S (Aux) OV] (17)a is dominant whereas the others are marginal. The primary frame associated with the proto-verb bijten is OV. The same holds for all other denotational verbs. One may interpret the appearance of the later construction (13)c as evidence that new exemplars with SVO structure are stored now, next to the analyses with pattern (13)b, which maintains its more than 2/3 dominance. If, by contrast, we assume that the earlier exemplars displaying the OV pattern ((13)a, (13)b) exert an overriding influence, the model needs a transformational move, modulo V<+finite>, to derive (13)c from (13)a and (13)b.

How to decide which of these conjectures is the right one? The answer comes from the next step in the acquisition process: (13d). In Dutch subordinate clauses (als Minou de muis bijt), the verb appears at the end. The subordinate clause is thus markedly different from the main clause, while it closely resembles the sentence of early child Dutch. Above we observed that Dutch children acquire the subordinate clause word order extremely fast, on the basis of very sparse input.

An elegant explanation in terms of a DOP-model would be, that the exemplars displaying the early child language word order are still accessible in the corpus. The underlying structures which are assumed in transformational grammar, would then be available due to earlier acquisition steps. The early exemplars thus play the same role as fixed theta frames (Brent 1994) in conventional grammars.

However, this account should also be applicable to verbs that the child encounters in main-clause exemplars, after it has acquired the verb-second word order. For that to be the case, there must be a cognitive relation between verb-second main clauses and their verb-final infinitival counterparts. This can be done in several ways.

Scha (1990) suggested the use of transformations in the style of Chomskyan grammars. For the example in (17)b this would mean that the model can construct the syntactic tree for Minou bijt de muis by applying the verb-second transformation to Minou de muis bijten. To achieve this, the generative machinery of the Stochastic Tree Substitution Grammar would have to be enriched with movement transformations.

An alternative possibility is suggested by the work on LFG-DOP, where the generative model consistently works with pairs of (sub-)trees: a surface (sub-)tree aligned with a tree that represents the "functional structure" of the sentence (cf. Bod & Kaplan 2003). In that approach, one should probably assume that in early child language surface structure and functional structure coincide. For the Dutch child, the acquisition of verb-second word order marks the moment that surface structure and functional structure split apart: the surface tree of a verb-second utterance is paired with the tree of the corresponding verb-final utterance to represent its functional structure.
4 Conclusion

There are developmental arguments for the thesis that Dutch is an SOV-language: SOV is the order of early Dutch child language, and the steps of the language acquisition process can only be understood if we assume that the early SOV-frames remain available in the later stages.

Exemplar-based language models may be able to explain this process, if they allow early exemplars to remain accessible long enough. Models that are exclusively based on tree-substitution, however, do not seem very plausible, since they cannot connect Dutch main clause word order to the underlying SOV pattern.

References


Abstract

Two studies are presented that show that exemplar frequency and similarity interacts with syntactic priming. In naturally-occurring spoken corpus data from the dative alternation and relative clause attachment, the strength of the priming effect is shown to depend on the exemplar frequency of the prime structure. Target exemplars that are more frequent are also more likely to occur. Also, using a precise definition of exemplars and exemplar similarity, more similar prime and target exemplars are shown to be more likely to prime. This is evidence for an exemplar model of the mental representation of syntactic knowledge, because the exemplar hypothesis predicts that exemplar similarity and the frequency of occurrence of exemplars will affect language processing.

1 Introduction

Exemplar models of language hold that entire strings of words are stored with their lexical and even phonetic content. Most of the evidence so far for exemplar models has come from phonology and morphology (Bybee, 2000; Bybee and McClelland, 2005; Krott et al., 2006), with the work on syntax consisting primarily of computational models (Bod, 2006), although Hay and Bresnan (2006) is an exception. The studies presented here provide empirical evidence for an exemplar model of syntactic representation by demonstrating that exemplar frequency and similarity affects syntactic priming.

Syntactic priming (Bock, 1986; Bock and Griffin, 2000) is the tendency for speakers to produce a syntactic construction if they have recently been exposed to that construction. Syntactic priming is interesting because it can be used to probe the nature of the mental representation of linguistic structures. As Branigan and Pickering argue (1995), priming results can be understood as a tendency to repeat structures (produce a "target") that has been heard or comprehended (the "prime") when the prime and target are similar along some dimension. If speakers’ behavior is sensitive to this similarity, then that similarity must arise from the two structures having the same cognitive representation of that dimension. Thus, by experimentally finding the dimensions of similarity between structures that cause priming, one can determine the nature of the mental representations of those structures. Applying this argument to the exemplar model of syntactic representation, exemplar models provide other dimensions of cognitive similarity for priming, and these can be tested by calculating exemplar frequency and similarity. If two exemplars are similar along their dimensions of representation, then one should be more likely to prime the other. The key prediction of exemplar models is that the exemplars must be compared along dimensions of very fine-grained detail. If such a similarity metric is found to influence language production, then that is a strong argument for the exemplar hypothesis. Similarly, the frequency of those exemplars should affect their likelihood of priming and production. Thus, if a speaker’s knowledge of syntax consists of stored fragments of language, then syntactic production, and syntactic priming, should be influenced by the frequency and similarity of these string tokens. Models of representation that do not include exemplar storage do not predict this interaction of priming with exemplar frequency and similarity.

In order to investigate whether priming is af-
fected by exemplar frequency and similarity, two corpus studies of English will be presented. These studies show priming is very sensitive to frequency and similarity effects. This is important empirical evidence that the mental representation of syntactic knowledge consists of exemplars. First, I will discuss a background study that was the first to show how syntactic priming is affected by the frequency of the prime, using data from the ditransitive alternation. Next, I present a new study showing frequency effects on priming of relative clause attachment. In a second study, I show that a precise metric of exemplar similarity, the nearest-neighbor distance, predicts the likelihood of priming. Finally, I conclude and discuss future work.

2 Background

Turning first to frequency effects on priming, the prediction of the exemplar model for the particular direction of the interaction between priming and frequency is not obvious. For the target, the direction of the effect of exemplar frequency is clear, more frequent exemplars should be more likely to be produced. Therefore, a positive effect on target production is predicted for target frequency. However, the direction of the effect of prime frequency is more interesting. It has been noted (Bock, 1986) that syntactic priming is sensitive to the frequency of the prime construction, with infrequent constructions priming more. This makes sense if priming is implicit learning (Bock and Griffin, 2000; Jaeger and Snider, 2007), which predicts that less frequent events lead to more activation (and hence more learning). This predicts a negative effect of exemplar frequency on priming strength.

In the first study to show this negative effect of prime frequency on the likelihood of priming, Jaeger and Snider (Jaeger and Snider, 2007) examined data from the ditransitive alternation, which is exemplified as follows:

(1)a. (NP PP) Yeah, I haven’t given much thought to it, I’m kind of busy raising my kids (SWBD)

b. (NP NP) Yeah, I haven’t given it much thought, I’m kind of busy raising my kids

They used an existing database of ditransitives from spontaneous speech (Bresnan et al., 2007), and extended the best currently available model of the choice between NP PP and NP NP (Bresnan et al., 2007) to contain additional controls and what I argue is a partial measure of exemplar frequency (the prime verb’s subcategorization bias; see below).

Their data consisted of the database compiled by Bresnan and colleagues (2007)1, which contains all 1,108 ditransitives with preceding primes from the full Switchboard corpus of approximately 2 million words (Godfrey et al., 1992). Bresnan and colleagues tested the effects of many factors on alternation choice in the ditransitive, including the role of syntactic persistence. They found that speakers are more likely to produce an NP PP structure like (1a) (the target) if the most recent ditransitive structure (the prime) was an NP PP structure (Recchia, 2007).

Similar to the two new studies presented later, Jaeger and Snider analyzed the data using a multiple logistic regression model. To ensure that their priming effects were not artifacts of other features of the local discourse environment, their model contained as controls all factors Bresnan and colleagues found significant in predicting the dative alternation (for a formal introduction to logistic regression, see Agresti (2002); for an informal introduction, see Jaeger (2007); see also below). They excluded the standard syntactic persistence factor used in the models of Bresnan and colleagues, instead, splitting the data set into two parts: one data set with all NP PP primes, and one with all NP NP primes. This was done because the effect of the prime’s frequency differs depending on the prime structure. Their study did not assume an exemplar model of representation, but they added a factor of frequency of the head verb in each possible ditransitive alternation, and this can be interpreted as a partial measure of exemplar frequency. The verb bias is defined as the conditional probability of the NP PP structure given the verb. As an example, take a verb like cost, which is highly biased towards NP NP:

(2)a. “A hard disk drive would cost several thousand dollars to the consumer…”2

b. “...inaccurate credit information could cost the consumer tens-of-thousands of dollars…”3

1The data set is available in the languageR package for the R statistical language
The verb *cost* is very rarely used in the NP PP structure, but as (2a) attests, it does occur. The prediction of the exemplar hypothesis is that the likelihood of a structure being repeated is affected by how likely the prime structure was to have occurred.

Here and in my studies following, I describe the coefficient for each independent variable and its levels of significance. Coefficients in logistic regression models are given in log-odds (the space in which logistic models are fitted to the data). For categorical factors, significant positive coefficients mean that a correct answer is more likely in the tested level of the variable than in the other level. For example, if the coefficient of Pronominal Theme for ditransitives is positive, then having a theme that is a pronoun makes the NP PP structure more likely. Negative coefficients mean the opposite. If the coefficient of Animate Recipient is negative, then having a recipient that is animate makes the NP PP structure less likely. For continuous factors, significant positive coefficients indicate how much more likely a correct answer is for each 1 unit increase in the level of the variable. For example if the coefficient for Length of Recipient is 0.72, then for each one word increase in the length of the recipient NP, the NP PP structure is \( e^{0.72} = 2.1 \) times more likely. In my later studies, I also report the difference in odds between conditions (as the name suggests, odds are simply \( e^{\text{log-odds}} \)). Odds range from 0 (for proportions of 0) to positive infinity (for proportions of 1), with proportions of 0.5 corresponding to odds of 1. Odds are a multiplicative scale, so I talk about an \( x \)-fold increase or decrease in odds between conditions.

Jaeger and Snider found that the prime’s verb bias (exemplar frequency) is a significant predictor of the NP PP structure \( (p < 0.05) \), with a negative log-odds coefficient. Thus, NP PP primes containing verbs that are biased against the NP PP structure make it more likely that the target will be NP PP than those containing NP NP-biased verbs. They also found a positive effect of target verb bias, or exemplar frequency \( (p < 0.001) \), with a positive log-odds coefficient, so verbs that more frequently occur in the NP PP make the NP PP structure more likely. In NP NP-prime data set, they also found a positive effect of target frequency \( (p < 0.001) \), but no effect of prime frequency.

Jaeger and Snider’s study showed that priming is sensitive to the verb biases of both the prime and target. It also shows that the direction of the effect of prime probability is negative, as less likely primes are more likely to be repeated. These effects of verb bias in the prime and target could be naturally explained by an exemplar model of syntactic representation, where the representations that produce the priming effect essentially contain lexical detail about the prime and target constructions. However, one could also argue that these results could be explained by the storage of frequencies in an abstract representation of argument structure, since the effects involve the biases of head verbs in particular argument structures. Thus I undertook the Study 1 to demonstrate the effect of exemplar frequency in priming non-argument constructions.

3 Relative clause attachment height (Study 1)

In this study, I examine exemplar frequency effects on priming in a structure that does not involve an argument relation: relative clauses (RCs), which are not usually considered to be arguments of their head nouns. Specifically, I investigate the priming of RC attachment in multiply-embedded NPs. As was first shown in an experiment by Scheepers (2003), subjects are more likely to complete a sentence with an RC attached “high” (to the first NP in an embedded NP structure like in (3a)) when they have previously processed such a high attachment, than when they had previously processed a “low” attachment (to the second NP).

(3a). (HI) I would definitely consider a try to find [\text{a school within [\text{the state}]} [\text{that I liked well enough to attend}]] (SWBD)

b. (LO) I would definitely consider a try to find [\text{a school within [\text{the state}]} [\text{that I live in}]]

3.1 Data

To test for priming and frequency effects in this structure, all the instances of two embedded NPs modified by an RC were extracted from the Penn Treebank Switchboard corpus (Marcus et al., 1994), a corpus of naturalistic, conversational data. There are a total of 610 prime-target pairs, but I excluded those for which there was no prime and for which I did not have a probability estimate.
(see below), leaving 272 prime-target pairs (190 LO-attached targets and 82 HI-attached targets).

3.2 Method

I again analyzed the data with multiple logistic regression models, with the HI attachment as the positive response. Because priming of this construction in naturalistic conversation has not been reported previously, a control study was undertaken, to determine if the prime RC attachment height affects the target attachment height. This was indeed the case, for a priming effect was found such that having a preceding HI attachment in the discourse makes a HI attachment much more likely in the next instance \( p < 0.001 \).

Having established priming for this construction in a corpus, exemplar frequency was operationalized as the frequency with which each of the two head nouns (school and state in the example above) occur with an RC, or the probability of the RC given the head noun. This probability was shown to be a predictor of relativizer (that) absence in Jaeger (2006), and the probabilities derived in that study were generously provided by Florian Jaeger for use in this one. Jaeger’s study actually examined non-subject extracted relative clauses (NSRCs), so the probability measure I use is actually \( P(\text{NSRC} \mid \text{head N}) \), which I consider to be an estimate of \( P(\text{RC} \mid \text{head N}) \). Since the previous study, and other studies in Jaeger and Snider (2007), showed that there is an inverse frequency effect in priming, a prime exemplar frequency factor was added that reflects this.

In order to determine how to calculate exemplar based expectation in a structure such as multiply-embedded NPs, which essentially have two heads to which an RC could attach, one needs to consider what would make a prime unexpected, or surprising, in the information theoretic sense of surprisal \( \text{surprisal}(X) = \frac{1}{\text{probability}(X)} \). If the prime is actually observed with a HI attached RC, then that prime would be considered unexpected if the LO NP had a high probability of being modified by an RC. Similarly, if the prime structure had a LO attached RC, then that would be unexpected if the HI NP had a high probability of being modified by an RC. Therefore, the prime exemplar surprisal factor was added as the probability of the non-modified noun in the prime to take an RC, thus if the prime had a HI attachment, the prime probability was the probability of the LO noun to have an RC, and if the prime had a LO attachment, the prime probability was the negative probability of the HI noun to have an RC. The factor is negative for LO attached primes, because these are expected to make the a HI attachment in the target less likely. The prime probability factor is defined in equation (1). This factor reflects the prediction that unexpected primes have a stronger effect: if the LO noun of the prime has a high likelihood of being modified by an RC, and the HI noun was actually the one modified by the RC, then this will increase the likelihood that the target will be a HI attachment.

To examine the exemplar frequency effects in the target, I added the probability of the HI and LO target nouns to be modified by an RC. As a final control, I added the distance in utterances between the prime and target.

3.3 Prediction

The exemplar hypothesis predicts that the exemplar frequency of the RC should affect processing. In this case, it predicts that the frequency of the prime RC (the probability that an RC would attach to its head noun in the prime structure) will affect the likelihood of priming. I expect that the probabilistic surprisal of an RC given the non-modified noun in the prime will interact positively with priming. The hypothesis also predicts that the probability of the target nouns to be modified by an RC will affect their likelihood of being modified in the target: the more likely an RC given the HI target noun, the more likely the target will be a HI attachment (a positive coefficient in the model), and the more likely an RC given the LO target noun, the more likely a LO attachment (a negative coefficient).

3.4 Results

The results of my models of priming of RC attachment are in Table 1. The coefficients in log-odds and standard errors associated with the remaining factors are given in the second and third column of Table 1. The corresponding odds coefficients are given the fourth columns. The fifth and sixth columns summarize the Wald’s Z statistic, which tests whether the coefficients are significantly different from zero (given the estimated standard error). Finally, the last two columns give the \( \chi^2 \) over the change in data likelihood \( (\Delta_x(\Lambda)) \) associated with the removal of the predictor \( (x) \) from the final model. The latter test is more robust against collinearity in the model (Agresti, 2002). The \( \chi^2 \)
value, which literally corresponds to the difference in the model’s data likelihood without the predictor, can be seen as a measure of the predictor's importance in the model. The Wald-test is included because it implicitly test the directionality of the effect (unlike the $\chi^2$ over the change in data likelihood). As the table shows, a main effect of prime surprisal was found ($p < .005$) such that less expected primes given the head nouns prime more. This is graphically represented in Figure 1. Interestingly, when the prime probability effect is controlled, there is no main effect of the type of prime structure (HI or LO). Fast backwards model comparison tests showed that the prime surprisal factor explains significantly more of the variance and therefore is the only significant effect. There are also significant effects of the probability of an RC given the the target nouns: if the HI target noun is frequently modified by an RC, then a HI attachment in the target is more likely ($p < .005$), also if the LO target noun is frequently modified by an RC, then a HI attachment in the target is less likely ($p < .005$). As in the previous study, there was no effect of prime-target distance.

3.5 Discussion

This study shows novel evidence for priming of RC attachments in naturalistic, conversational corpus data. It also provides further evidence for the exemplar hypothesis: the priming effect is sensitive to the exemplar frequencies of both the prime and target. This further strengthens the argument for the exemplar hypothesis because the study involved a non-argument construction and therefore could not be due to abstract representations of argument structure.

4 Exemplar similarity in ditransitives (Study 2)

The above results argue for an effect of exemplar frequency on priming. It must be acknowledged, however, that the particular results presented here could be explained by other models of representation. For example a lexicalized probabilistic context free grammar, where syntactic knowledge consists of rules and their probabilities given their head word, provides representations that could explain these results. Also, I have not presented a clear model of the exemplars. In this study, I remedy both of these problems by testing exemplar priming effects in a model like that used in $k$-nearest-neighbor ($k$-NN) classifiers (Aha et al., 1991; Daelemans and van den Bosch, 2005).

$k$-NN models classify cases by comparing the

$$\text{prime surprisal} = \begin{cases} P(RC|\text{LO noun}) & \text{if prime = HI} \\ -P(RC|\text{HI noun}) & \text{if prime = LO} \end{cases}$$ (1)
current case with all other observed exemplars (usually a data set derived from a corpus), and picking the best classification based on past experience. An essential part of this model is the distance metric used to determine how similar exemplars are to one another. The algorithm uses this to weight highly the exemplars that are closest to the case to be classified. However, such a distance metric could also be extremely useful in determining whether one structure can prime another. Recall that Branigan and Pickering (1995) argued that priming occurs when the prime and target are similar along some dimension. The k-NN model of exemplar similarity provides a metric for explicitly measuring the similarity of the prime and target. If exemplars are the correct representation, then this measure of exemplar similarity should predict the likelihood of priming.

Exemplars in k-NN models are defined as sets of features. The distance between exemplar X and exemplar Y, \( \Delta(X, Y) \), is the weighted sum of the difference between the two exemplars along each of the dimensions defined by the features:

\[
\Delta(X, Y) = \sum_i w_i \delta(x_i, y_i)
\]

where \( w_i \) is the weight associated with feature \( i \), and \( \delta(x_i, y_i) \) is defined as:

\[
\delta(x_i, y_i) = \begin{cases} 
\frac{x_i - y_i}{\text{max} - \text{min}} & \text{if numeric, otherwise} \\
0 & \text{if } x_i = y_i \\
1 & \text{if } x_i \neq y_i
\end{cases}
\]

If the feature is continuous, then the difference between the two exemplars is just the difference in the values of feature \( i \) for the two exemplars, divided by the maximum range of that feature (so that features won’t be weighted too highly just because they involve larger numbers). If the feature is categorical, the the difference between the two exemplars is 0 if they have the same value for the feature, and 1 if they have different values for the feature.

Applying such a model to priming, the basic idea is that two exemplars that are more similar, or have a lower NN-distance between them, are more likely to prime. This study tests this prediction in the ditransitive alternation.

4.1 Data
I use the same data as in Bresnan and colleagues (2007) and Jaeger and Snider (2007). For greater statistical power, I use the full data set of 1,108 ditransitives with preceding primes (as opposed to splitting it in two) from the full Switchboard corpus. Again, the data consists of naturalistic conversation.

4.2 Method
I again analyzed the data with logistic regression models. I added to the ditransitive data set an exemplar NN-distance metric as defined in equation (2). The similarity space in which the distance was calculated is quite fine-grained, consistent with an exemplar model of representation. It has as dimensions all the features used to predict the dative alternation in the original study by Bresnan and colleagues (2007). Thus, in the ditransitive data set, the prime and target exemplars can be compared in terms of the pronominality, animacy, givenness, etc. of the recipient and theme arguments, as well as the verb identity (Bresnan and colleagues modeled this as a random effect). Verb identity between prime and target has been found to correlate positively with priming (Pickering and Branigan, 1998), which is an obvious prediction of the exemplar similarity hypothesis: exemplars with the same verb are clearly more similar.

In order to determine the feature weights used in equation (2). I ran the k-NN classifier TiMBL (Daelemans and van den Bosch, 2005) on the entire ditransitive data set, using the same features as Bresnan and colleagues. Using Information Gain weighting, TiMBL classified the ditransitive alternation with 92.8% accuracy, which is comparable to the 96% originally found by Bresnan and colleagues. Using the feature weights determined by TiMBL, I calculated the exemplar NN-distance between each prime and target in the ditransitive data set.

4.3 Prediction
The NN-distance metric can be used to define a similarity metric between the prime and target exemplars in a data set. The prediction is that the lower the distance between the prime and target exemplars, the more likely the prime and target will occur in the same construction.

4.4 Results
The results of the ditransitive analysis are in Table 2. The first 14 rows in Table 2 summarize controls from Bresnan et al and Snider and Jaeger. The last
Table 2: Summary of ditransitive analysis

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Parameter estimates</th>
<th>Wald’s test</th>
<th>Δχ²(Λ)-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(independent variable)</td>
<td>Log-odds</td>
<td>S.E.</td>
<td>Odds</td>
</tr>
<tr>
<td>verb class = communication</td>
<td>-2.505</td>
<td>0.631</td>
<td>0.06</td>
</tr>
<tr>
<td>verb class = future transfer</td>
<td>-0.154</td>
<td>0.812</td>
<td>0.94</td>
</tr>
<tr>
<td>verb class = prevent transfer</td>
<td>-4.598</td>
<td>2.409</td>
<td>0.01</td>
</tr>
<tr>
<td>verb class = transfer</td>
<td>0.434</td>
<td>0.418</td>
<td>1.22</td>
</tr>
<tr>
<td>recipient nongiven</td>
<td>2.488</td>
<td>0.462</td>
<td>10.69</td>
</tr>
<tr>
<td>theme nongiven</td>
<td>-1.467</td>
<td>0.442</td>
<td>0.22</td>
</tr>
<tr>
<td>recipient pronominal</td>
<td>-0.288</td>
<td>0.484</td>
<td>0.73</td>
</tr>
<tr>
<td>theme pronominal</td>
<td>1.469</td>
<td>0.449</td>
<td>4.06</td>
</tr>
<tr>
<td>theme indefinite</td>
<td>-2.146</td>
<td>0.411</td>
<td>0.11</td>
</tr>
<tr>
<td>recipient inanimate</td>
<td>3.481</td>
<td>0.579</td>
<td>28.51</td>
</tr>
<tr>
<td>recipient non-local person</td>
<td>0.676</td>
<td>0.390</td>
<td>2.02</td>
</tr>
<tr>
<td>theme singular</td>
<td>-0.995</td>
<td>0.377</td>
<td>0.43</td>
</tr>
<tr>
<td>log argument length difference</td>
<td>-1.342</td>
<td>0.243</td>
<td>0.24</td>
</tr>
<tr>
<td>target verb bias</td>
<td>3.710</td>
<td>0.730</td>
<td>42.34</td>
</tr>
<tr>
<td>prime=NPPP</td>
<td>1.665</td>
<td>1.236</td>
<td>11.85</td>
</tr>
<tr>
<td>prime V = target V * prime=NPPP</td>
<td>-0.024</td>
<td>0.825</td>
<td>0.97</td>
</tr>
<tr>
<td>NN distance * prime=NPPP</td>
<td>-2.365</td>
<td>0.869</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Row (17) contains the result of interest, that exemplar similarity is a significant predictor of priming. The negative coefficient indicates that, if the prime is NP PP, and the NN-distance between the prime and target is small, then the target is more likely to be NP PP. This is graphically represented in Figure 2. This effect is still significant controlling for prime-target verb identity, because row 16 shows that factor is not significant. This indicates that the exemplar similarity effect is not due just to the verb identity effect of Pickering and Branigan (1998). Row 15 shows the priming due to the previous structure being NP PP is not significant as a main effect, but its interactions are significant, as rows 16 and 17 show. Finally, although it is not shown in the table, I tested for the effect of prime surprisal. I found it to be marginally significant (p < 0.1). This is probably due to the null effect that Jaeger and Snider found when the prime structure is NP NP.

4.5 Discussion

This study shows that priming is sensitive to exemplar similarity. The target is more likely to occur in same structure as the prime when the NN-distance between the prime and target is small. Further, the exemplar similarity metric used is quite precisely defined and provides and clear model of the exemplar representation. Also, the effect of exemplar similarity is not likely to be an artifact of local discourse and syntactic factors, because these were controlled.

5 Conclusions and Future Work

In conclusion, these two studies have shown exemplar frequency and similarity interacts with priming. Specifically, the strength of the priming effect depends of the exemplar frequency of the prime
structure. Also, target structures that are more frequent are also more likely to be produced. This is evidence for an exemplar model of the mental representation of syntactic knowledge, because the exemplar hypothesis predicts that the frequency of occurrence of exemplars will affect language processing. Further, priming was shown to be more likely when the prime and target are similar. This is particularly strong evidence for the exemplar hypothesis because a precise model was used to define the exemplars and their similarity metric.

In future work, I will further refine the tests of the exemplar hypothesis. I plan to add new features to the distance metric that represent the degree of similarity between other words in the prime and target structures, for example the head nouns of the recipient and theme arguments. This is a particularly important test of the exemplar hypothesis, because the hypothesis predicts that the representations contain such fine detail as the head words involved, in addition to the more general features already in the data set.

6 Acknowledgments

I am especially grateful to Florian Jaeger for advice on analysis and on describing the modeling techniques. I also thank Joan Bresnan for discussions and advice on interpretation. Thanks to Joan Bresnan, Gabe Recchia, and colleagues for providing me with the database of ditransitive prime-target pairs from the Switchboard corpus. Also, thanks to Tom Wasow for very helpful feedback and questions.

References


T. Florian Jaeger. 2007. Better categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. submitted.


Abstract

Data Oriented Parsing (DOP) is an exemplar-based model of language use that processes new input based on past experience by combining structural fragments extracted from a given treebank. In the simplest case (Tree-DOP) these fragments are subparts of simple phrase structure trees, each associated with some probability. The approach is attractive in many ways but the impoverished representational basis is a serious drawback from a linguistic point of view. This paper describes the theoretical characteristics of a linguistically richer version of DOP based on the Head-driven Phrase Structure Grammar (HPSG) formalism.

1 Introduction

Evidence of the probabilistic properties displayed in human language processing (Juliano and Tanenhaus, 1993; Jurafsky, 1996) has led to the statistical enrichment of Natural Language Processing (NLP) models. One approach to this involves associating the rules of a competence grammar with probabilities computed from large-scale syntactically annotated corpora. Simply adding probabilities to rules, however, is of limited value because disambiguation preferences are “memory-based” and can depend on arbitrarily large syntactic constructions (Bod, 2003). Exemplar-based models of language use have gained ground in recent research.

A well-known model based on such an approach is Data Oriented Parsing (DOP) (Bod, 1992; Bod, 1995) which processes new input by combining fragments extracted from a given treebank. In the simplest case (Tree-DOP) these fragments are subparts of simple phrase structure trees produced by two decomposition operations; Root and Frontier. Root creates passive fragments by extracting substructures as in Figure 1(b), while Frontier produces active fragments by deleting pieces of substructure as in Figure 1(c). Each fragment is assigned some probability mass based on a predefined estimator. Disambiguation involves finding the structure(s) with the highest probability.

![Figure 1: Decomposition of Jane runs in Tree-DOP](image)

The approach is attractive in many ways but the impoverished representational basis is a serious drawback from a linguistic point of view. Bod and Kaplan (1998) address this issue by proposing a linguistically richer version of DOP based on the more sophisticated Lexical Functional Grammar (LFG) representations. The resulting model (LFG-DOP)
constitutes a very powerful model of language performance, but it also suffers from several disadvantages. The first of these relates to the degree of generality of the fragments produced. The standard decomposition operations create fragments that are over-specific, leading to under-generation and exacerbating the normal problem of data sparsity. Root and Frontier, for example, produce fragments like (1) \( a \) and \( b \) from the corpus representation of “Jane runs”. These fragments, however, will not be usable in parsing either “Sam likes Jane”, because Jane is not acc, or “Jack runs”, assuming Jack is not fem.

\[
\begin{align*}
(1) & \quad a. [NP \text{Jane}]_{3rd/sg/fem/nom} \\
& \quad b. [s \text{NP}_{3rd/sg/fem/nom} \text{ runs} ]
\end{align*}
\]

To overcome this, Bod and Kaplan formulate a third decomposition operation known as Discard which generalises over the fragments produced by the other two. Discard, however, applies in a highly unconstrained manner (i) causing the size of the fragment corpus to explode and (ii) allowing under-specific fragments like (2) to be produced. The latter leads to overgeneration problems (e.g. *Him runs). To deal with this, Bod and Kaplan redefine ‘grammaticality’ by ruling out analyses that cannot be derived without fragments produced by Discard. The corpus-based dimension of this definition, however, cannot extend it beyond the training data. Suppose, for example, that a particular noun only appears in the treebank as the head of object NPs (where it is marked accusative). Sentences where such a noun appears as head of a subject NP, where it must be nominative, will only be derivable using Discard fragments, and will hence be considered ungrammatical.

\[
(2) \quad [s \text{NP}_{3rd/sg} \text{ runs} ]
\]

In addition, not all of LFG’s well-formedness conditions can be checked efficiently during the derivation process, resulting in some probability mass being assigned to invalid structures. This probability mass is hence “wasted” raising theoretical questions about the use of relative frequency estimation in computing the most likely analysis.

In the following section we set the theoretical background of a novel version of DOP based on Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994), that addresses these issues. Previous attempts to define such a model (Neumann, 1999; Neumann, 2003) were based on extracting a Stochastic Lexicalised Tree Grammar (SLTG) from an HPSG parsed training corpus and using it in a manner similar to Tree-DOP. Node labels in the trees represent the HPSG rule-schema applied during the corresponding derivation step. A complete parse tree can be unfolded into an HPSG representation by expanding the rule labels and lexical types to the corresponding feature structures. Despite their differences, this approach suffers in some cases from the same problems as LFG-DOP. Well-formedness of the resulting structure, for example, cannot be ensured in an efficient manner because some of the parse trees produced cannot be unfolded into valid feature structures. This approach, however, does not provide the only, or even the most linguistically well-founded way of moving towards HPSG-DOP.

2 HPSG-DOP

Presenting a DOP model involves instantiating the following four parameters: (i) how utterances are represented; (ii) how representations are decomposed into fragments; (iii) how fragments are combined; and (iv) how the proposed analyses are disambiguated.

2.1 Representation

The representational framework we assume is conventional HPSG, along the lines of Ginzburg and Sag (2000). The HPSG linguistic ontology is a system of signs. These can be either of type phrase describing phrasal constituents or lex-sign describing words and lexemes (3). All signs have the top level attributes PHON and SYNSEM which describe the phonological content and the syntactico-semantic characteristics of the sign in question respectively. In addition, signs of type phrase carry the attribute DTRS (daughters) describing the surface constituency of the phrase. Lexical signs, on the other hand, possess the feature ARG-ST whose value is an ordered list of objects corresponding to the arguments (subject, specifier and any complements) required by the lex-sign being described.
We will draw feature structures either as Directed Acyclic Graphs (DAGs) or as Attribute-Value Matrices (AVMs), using a wide range of abbreviations (e.g. ‘NP’ stands for a nominal phrase with empty SPR and COMPS lists, ‘VP’ stands for a verb-headed phrase that subcategorises for a subject, while ‘S’ stands for a verb-headed phrase with an empty SUBJ list). Figure 2 gives the DAG representation of the sentence “Jane runs”, while Figure 3 presents it as an AVM (both somewhat simplified).

We assume representations are totally well-typed feature structures. A representation is totally well-typed if all and only the required attributes are present and each of them has an appropriate value. Of course, this only makes sense against the background of a particular type theory, that is, a signature, which defines an hierarchy of types, and a collection of type constraints which indicate, inter alia, what combinations of attributes and values are permitted for different types. Fragments should respect the same principles as the representations they are produced from: i.e. they should be totally well-typed feature structures. The total well-typedness requirement implies that fragments may be subject to a form of type inference which we will refer to as type expansion:

**Definition 2.1 (TypeExp)** Let $F$ be a feature structure, and $T$ a type theory, then $\text{TypeExp}_T(F)$ is the most general totally well-typed extension of $F$ according to $T$ such that $F \subseteq \text{TypeExp}_T(F)$.

Type expanding the sort $\text{hd-str}$, for example, produces the feature structure in Figure 4 (assuming the type theory in (3)).

---

**Figure 1:** DAG representation of Jane runs.

**Figure 2:** DAG representation of Jane runs.

**Figure 3:** AVM representation of Jane runs.

**Figure 4:** $\text{TypeExp}(\text{hd-str})$
Type expansion can produce very specific results. In addition to adding attribute-value pairs, the constraints on certain types can enforce re-entrancies between various parts of the feature structure (e.g. the type hd-subj-str requires that the SUBJ value of its head-daughter and the SYNSEMs value of its non-head daughter are structure-shared). This is due to the fact that essentially the whole grammar, including the lexicon, is expressed as a collection of type constraints on a signature.

### 2.2 Decomposition Operations

Decomposition in HPSG-DOP is carried out by Root and Frontier. Before extending these operations so that they become applicable to feature structures, we will introduce some terminology. Let the descendants (of a sign), be recursively defined as the elements of the sign’s DTRS list, and their descendants. In addition, let $F$ be a feature structure with a descendant $D_F$. Suppose $D_F$ is removed from $F$ giving rise to $F'$. Then $Context(D_F)$ denotes the subgraph rooted at the removal node in $TypeExp_T(F')$, and $Inherent(D_F)$ denotes the set-theoretic relative complement of $D_F$ and $Context(D_F)$. An example will clarify this.

Suppose $F$ is the feature structure in Figure 2 and $D_F$ is its n-lx daughter. If $D_F$ is removed from $F$ it gives rise to a structure $F'$ like the one in Figure 5. On standard assumptions, $F'$ is type-expanded to $TypeExp(F')$ as in Figure 6.

These standard assumptions require that what fills the $DTRS|1ST$ slot be an object of type sign, whose PHON is re-entrant with the first part of the PHON of the whole sentence (i.e. tag $\downarrow 1\uparrow$ which follows from general constraints on $hd-subj-str$). In addition, the sign’s SYNSEM (SS) value will be re-entrant with the the $SUBJ|1ST$ slot of $runs$, which restricts it to being a 3rd person, singular, nominative nominal. Notice, there will be no constraint that requires the subject of $runs$ to be $fem$. $Context(D_F)$ then denotes the subgraph rooted at $sign^*$ in Figure 6, which is roughly $NP_{nom.3.sg}$.

Intuitively, these feature-value pairs could result from $Jane$ being in that particular context (e.g. it might be that $3.sg$ results from $Jane$ being the subject of $runs$). $Inherent(D_F)$ denotes the relative
complement of $D_F$ and Context ($D_F$) as depicted in Figure 7. It includes features that cannot reside in the context such as inherent phonological and semantic features of the entity in question, notably that its phonological content is /Jane/ and that it is feminine.

![Figure 7: Inherent ($D_F$)](image)

We will now turn to defining the decomposition operations, starting with Root.

**Definition 2.2 (Root)** Given a representation $F$ licensed by a type theory $T$, Root selects any descendant $D_F$ of $F$ and returns $\text{TypeExp}_T(\text{Inherent} (D_F))$.

Suppose, for example, Root applies to Jane (i.e. the value of $DTRS|IST$) in Figure 2. It will return $\text{TypeExp}_T(\text{Inherent} (\text{Jane}))$, the type-expansion of the structure in Figure 7 (depicted in Figure 8) which has the properties of being nominal, and 3rd person singular but not nominative. The case restriction does not form part of $\text{TypeExp}_T(\text{Inherent} (\text{Jane}))$ since there is nothing in the type theory to force it to ‘grow back’. Notice that this fragment is of the right level of generality. Unlike the corresponding LFG-DOP fragment for Jane in example (1)a, this is 3.sg.fem, but not nom.$^1$

![Figure 8: TypeExp$_T$(Inherent (Jane)).](image)

**Definition 2.3 (Frontier)** Frontier erases any combination of $F$’s descendant’s and type expands the result $F'$ marking the erasure points for composition.$^2$

If Frontier applies to Jane (i.e. the value of $DTRS|IST$) in Figure 2, it will first erase the sub-structure corresponding to Jane as in Figure 5 marking the erasure point with $*$ and then type-expand the result (Figure 6). Notice, Frontier produces again fragments of the right level of generality (i.e. general enough to allow both masculine and feminine subjects, but not sufficiently general to allow accusative subjects (e.g. *Him runs) as was the case for LFG-DOP in example (2)). Another fact about Root and Frontier as formulated here is that they do not require Discard to generalise over the fragments they produce, which is what causes the size of the fragment corpus to explode in the case of LFG-DOP.

### 2.3 Head-driven Composition

Standard composition approaches in DOP are rightwards or incrementally rightwards directed (Bod, 1995; Neumann, 2003). In the context of HPSG, however, it is interesting to consider a “head-driven” approach to composition, whereby it is the head chain of the derivation initial fragment that identifies the order in which expansion nodes are to be considered as composition sites. Starting from the bottom, the open slot nodes of an *active fragment* are unified with other fragments so that each node along the path leading from the head lexical anchor to the root of the feature structure dominates a *passive* sub-constituent before the next node along the path is considered. Computation is bidirectional with the direction being identified at each step (rather than in some predefined manner) by the head chain of the fragment rooted at the node being considered. Such a process is of course reminiscent of head-driven parsing strategies (Proudian and Pollard, 1985; van Noord, 1997, etc.).

$^1$Notice that had Root been applied to a node possessing the phonology *she* it would have produced something that is nom because of a type constraint in the lexicon requiring this.

$^2$This definition produces fragments analogous to those of Tree-DOP. Taking into account, however, that node labels in HPSG-DOP are subtypes of sign which do not convey any syntactic information, one might want to formulate Frontier so that it cannot apply to the overall lexical head restricting fragments to a minimum of one lexical anchor in order to maintain some syntactic information in the fragment (Linardakis, 2006).
Figure 9 shows an example of head-driven composition in deriving a representation of “She found him” using the feature structure fragments corresponding to the subtrees on the previous page. The first internal node along the head chain of the derivation initial fragment (i.e. $hd$-$comp$-$str$) dominates an active subconstituent. The rightmost terminal node is, therefore, the first node to be expanded. Unification proves successful and since $hd$-$comp$-$str$ now dominates a passive constituent the pointer is advanced one step along the path of the head chain to the root node $hd$-$subj$-$str$. This again dominates an active subconstituent, thus identifying the next composition site. Fragment unification is again successful producing the last feature structure in Figure 9. This representation is totally well-typed and is, therefore, valid.

2.4 Fragment Probabilities
As in other DOP models, an HPSG-DOP representation will typically have many different derivations, and any string many different representations. Assuming composition steps are independent events, the probabilities of a derivation $d_j = \langle f_1, \ldots, f_n \rangle$ and a final representation $R$ with $m$ derivations $d_j$ are defined as in (4) and (5) respectively.
In order to compute fragment probabilities, we will use the standard relative frequency estimator as a starting point. In Tree-DOP subtree probabilities are defined as in (6). The set of all composable subtrees at each derivation step is hence identified by its category root(f). Category matching in HPSG-DOP would correspond to classifying fragments based on their head features and subcategorisation frame. Two fragments would be considered as competing if they shared the same values for all CAT features (i.e. HEAD, SUBJ, SPR and COMPS). The probability of a fragment f would then be defined as in (7). Underspecified HEAD values would be expanded in all possible ways and the resulting fragments classified accordingly.

\[
P(f_i) = \frac{|f_i|}{\sum_{root(f)=root(f_i)} |f|}
\]

(6)

\[
P(f_i) = \frac{|f_i|}{\sum_{\forall(v(SS/LOC/CAT,r(f))=v(SS/LOC/CAT,r(f_i)))} |f|}
\]

(7)

This stochastic process, however, is not guaranteed to identify a probability distribution over the set of valid representations because the combinatorial potential of a fragment is not entirely determined by its CAT value. As a result, this process assigns some probability mass to structures outside the parse space (much like in LFG-DOP) causing it to be wasted.

Abney (1997) argues that relative frequency estimation constitutes a nonoptimal approach to probabilistic Attribute-Value Grammars (AVGs) in general due to the independence assumption not being applicable to deep linguistic analyses because their fragments are equipped to handle both syntactic and semantic dependencies. Loglinear or maximum entropy models (Abney, 1997; Miyao and Tsujii, 2002) are generally deemed as more suitable for such formalisms because they do not rely on the independence assumption.

The problem can be avoided by allowing competition sets to include all fragments that can be successfully unified with the next composition site (NCS) of some other fragment. Since previous derivation steps can affect the specificity of such sites, competition sets in HPSG-DOP cannot be predetermined. Suppose f(i-1) is the structure produced before the ith step of the derivation process. The probability of the next fragment to be used is defined as in (8). Fragment probabilities for the derivation initial selection can be based on category matching relative frequency estimation since there are no previous derivation steps to determine unifiability.

\[
P(f_i) = \frac{|f_i|}{\sum_{\forall(f) f \text{ is unifiable with } NCS(f(i-1))} |f|}
\]

(8)

3 Discussion

HPSG-DOP enjoys a number of positive characteristics. The most salient of these is its great linguistic sensitivity. HPSG-DOP’s linguistic power, however, relies on grammaticality being defined entirely in terms of the type theory. This is nearly, but not quite, true in standard HPSG. One example of information being determined outside the type theory is the nominal reference of exempt anaphors.

Principle A of the Binding theory states that a locally a-commanded anaphor must be locally a-bound.3 This implies that anaphors that are not locally a-commanded need not be a-bound. In (9) and (10), for example, himself and themselves are not a-commanded because the ARG-ST value of picture(s) contains only one element (i.e. a PP). Such anaphors are known as exempt because they are exempted from the binding conditions.

3 An object X is said to locally a-command another object Y, both appearing on the same ARG-ST value, iff X precedes (i.e. is to the left of) Y on the ARG-ST value. X is said to a-bind Y if it a-commands it and the INDX values of X and Y are structure shared.
(9)  \textit{John}_j \text{ took a picture of himself}_j .

(10)  \textit{They}_i \text{ saw pictures of themselves}_i .

The implication of “need not” is what makes identifying the nominal reference of such anaphors go astray, because it does not determine whether \textit{exempt anaphors} are, in fact, a-bound or not, and if yes to what. Consequently, the type theory in these cases licences more than what is intuitively correct. In the case of “\textit{John} took a picture of himself”, for example, “\textit{John}” does not have to be coindexed with “\textit{himself}”, so (11) is perfectly acceptable for the type theory and, for the same reason, so is (12).

(11)  *\textit{John}_i \text{ took a picture of himself}_j .

(12)  *\textit{Mary} \text{ took a picture of himself}.

Next we will examine the effect of this in the context of HPSG-DOP. Take a simple training corpus containing the representation of the sentence “\textit{John} \text{ took a picture of himself}” in (13).

(13)  \textit{hd-subj-str} / \textit{S} /  

\textit{n-lx} / \textit{NP} /  

\textit{hd-comp-str} / \textit{VP} /  

\textit{transy-lx} / \textit{N} /  

\textit{hd-spr-str} / \textit{NP} /  

\textit{took} / \textit{D} /  

\textit{hd-adj-str} / \textit{N} /  

\textit{a} / \textit{N} /  

\textit{hd-comp-str} / \textit{PP} /  

\textit{picture} / \textit{P} /  

\textit{of} / \textit{NP} /  

\textit{himself}_j

Applying \textit{Frontier} to the \textit{NP} _{\textit{John}} node will produce the structure in (14). The implications of this are quite serious. Even though the produced fragment is valid, recombining it with the fragment that was cut off (i.e. \textit{NP} _{\textit{John}}) will not result in the initial structure because there is no constraint in the signature to reenforce the reentrance. In fact the structure in (13) cannot be reproduced in any way.

(14)  \textit{hd-subj-str} / \textit{S} /  

\textit{sign} / \textit{NP} /  

\textit{hd-comp-str} / \textit{VP} /  

\textit{transy-lx} / \textit{N} /  

\textit{hd-spr-str} / \textit{NP} /  

\textit{took} / \textit{D} /  

\textit{hd-adj-str} / \textit{N} /  

\textit{a} / \textit{N} /  

\textit{hd-comp-str} / \textit{PP} /  

\textit{picture} / \textit{P} /  

\textit{of} / \textit{NP} /  

\textit{himself}_j

The source of the problem discussed here is that the type theory “overgenerates”. One way of looking at this situation is hence to adopt the HPSG point of view according to which the occurrence of \textit{a-free} anaphors is linguistically valid so long as they are exempt. Even though this argument may be defensible from a linguistic point of view, it constitutes a less than satisfactory solution from the data-oriented point of view. The fact that no matter how much we overtrain a fragment corpus we will never be able to capture these dependencies when analysing new input stands in sharp tension with the DOP philosophy.

4 Concluding Remarks

We have presented a DOP model based on the syntactically and semantically articulated representations of HPSG. Of course, the general architecture portrayed here allows for various HPSG-DOP instantiations that can differ in the degree of specificity the fragments are allowed to have and/or the stochastic process employed for disambiguation.

Apart from the advantages that follow from the richer representational basis of the model, HPSG-DOP has a number of attractions. It takes full advantage of the type theory, thus enabling fragments produced to capture dependencies beyond the syntactic level, while at the same time being of the right level of generality. Additionally, well-formedness of the representation being built is checked throughout the derivation process. The outcome of successful com-
position of totally well-typed feature structures will itself be at least well-typed and certainly expandable to a totally well-typed feature structure (i.e. a valid representation). As a result, relative frequency estimation can provide a feasible basis for computing the most probable analysis in HPSG-DOP, unlike other statistically enriched unification-based models.

HPSG-DOP’s linguistic power, however, relies on grammaticality being defined entirely in terms of the signature. Unfortunately, this is not always the case in standard HPSG, where a number of phenomena are described outside the signature. In such cases decomposing an initial representation and composing it again is not guaranteed to reproduce the same structure.

A natural objection from a language engineering point of view is that one of the attractions of the data-oriented philosophy (that it seems to dispense with the need of grammars) has been lost. Our approach lacks this attraction because it relies on the existence of a type system (i.e. an HPSG grammar). From a theoretical point of view, however, it is reasonable to have both a performance model (i.e. a form of DOP), and a competence grammar (i.e. a type theory).

Another positive data-oriented characteristic which has been sacrificed in order to ensure fragments are of the right level of generality is the feature of robustness (i.e. the ability to deal with input which is in some way ill-formed or extra-grammatical). This issue can be approached in a number of ways. In the case of unknown words, for example, the techniques described for Tree-DOP by (Bod, 1995) can be straightforwardly extended to this model. Robust unification (Fouvry, 2003), which is based on extending the signature to a lattice to include the unique joins of every set of incompatible types, provides a promising alternative to this issue. While we have discussed how HPSG-DOP behaves from a theoretical point of view, its empirical evaluation remains outstanding.

References


Analogical Modeling: An Update

David Eddington and Deryle Lonsdale
Department of Linguistics & English Language
Brigham Young University
Provo, UT, USA 84602
{eddington,lonz}@byu.edu

Abstract

Analogical modeling is a supervised exemplar-based approach that has been widely applied to predict linguistic behavior. The paradigm has been well documented in the linguistics and cognition literature, but is less well known to the machine learning community. This paper sets out some of the basics of the approach, including a simplified example of the fundamental algorithm’s operation. It then surveys some of the recent analogical modeling language applications, and sketches how the computational system has been enhanced lately to offer users increased flexibility and processing power. Some comparisons and contrasts are drawn between analogical modeling and other language modeling and machine learning approaches. The paper concludes with a discussion of ongoing issues that still confront developers and users of the analogical modeling framework.

1 Introduction

Analogical modeling (AM) is an exemplar-based modeling approach designed to predict linguistic behavior on the basis of stored memory tokens (Skousen, 1989; Skousen, 1992; Skousen, 1995; Skousen, 1998). It is founded on the premise that previous linguistic experience is stored in the mental lexicon. When the need arises to determine some linguistic behavior (pronunciation, morphological relationship, word, etc.), the lexicon itself is accessed. A search is conducted for the stored exemplars that are most similar to the one whose behavior is being predicted. The behavior of highly similar stored entities generally predicts the behavior of the one in question, although less similar ones have a small chance of applying as well.

AM has been implemented in a succession of computer implementations that allow users to specify an exemplar base and then to test unseen instances of language behavior against this accumulated store to predict (and quantify) the relevant and possible outcome(s).

A comprehensive examination of the system is beyond the scope of this paper1. Instead, an attempt will be made to situate AM within the space of exemplar-based modeling systems and, to a certain extent, within the larger context of machine learning systems and cognitive modeling systems in general.

This paper begins by giving an informal overview of the algorithm used to perform AM. To accomplish this it traces in schematic form an early but typical application of the AM approach: Spanish gender guessing. The subsequent section discusses AM with respect to other types of modeling systems, and gives an overview of the types of linguistic phenomena modeled by AM. The paper then discusses several items of recent and ongoing work with the AM program. The conclusion mentions possible areas for future improvement and investigation.

1For further information including a bibliography and software downloads see the project webpage at: http://humanities.byu.edu/am
2 The AM algorithm

Perhaps the best way to understand the AM algorithm is with a concrete illustration. Predictions are always made in terms of specific exemplars, therefore, for the purposes of the example, the following seven monosyllabic Spanish nouns and their corresponding gender will be considered:

<table>
<thead>
<tr>
<th>Subcontexts</th>
<th>Members of the Subcontext</th>
<th>Pointers</th>
<th># of Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>pan</td>
<td>none</td>
<td>none</td>
<td>0</td>
</tr>
<tr>
<td>¿pan</td>
<td>plan M</td>
<td>plan M &gt; plan M</td>
<td>0</td>
</tr>
<tr>
<td>ñan</td>
<td>none</td>
<td>none</td>
<td>0</td>
</tr>
<tr>
<td>ñañ</td>
<td>paz F, par M</td>
<td>paz F &gt; paz F</td>
<td>2</td>
</tr>
<tr>
<td>ñañ</td>
<td>cal F</td>
<td>cal F &gt; cal F</td>
<td>0</td>
</tr>
<tr>
<td>ñañ</td>
<td>tren M, ron M</td>
<td>tren M &gt; tren M</td>
<td>0</td>
</tr>
<tr>
<td>ñañ</td>
<td>none</td>
<td>none</td>
<td>0</td>
</tr>
<tr>
<td>ñañ</td>
<td>rey M</td>
<td>rey M &gt; rey M</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Subcontexts and their disagreements.

are marked with asterisks in Table 1. A disagreement occurs when words that are equally similar to the given context, exhibit different behaviors, in this case, different genders. The number of disagreements is determined by pairing all members of a subcontext with every other member, including itself, by means of unidirectional pointers, and counting the number of times the members of the pair have different behaviors. In this example, the only subcontext containing any disagreement is ñañ.

Subcontexts are then arranged into more comprehensive groups called supracontexts as shown in Table 2, where a hyphen indicates a wildcard.

In the subcontextual analysis a tally of the all of the subcontextual disagreements is made. The supracontextual analysis consists of analyzing all of the words that appear in a given supracontext, and again tallying disagreements (Table 3).

In the supracontextual analysis, words that have more than one variable in common with pan appear in more than one supracontext. This is AM’s way of allowing the gender of words which are more similar to pan to influence the gender assignment of pan to a greater extent.

The purpose of AM’s algorithm is to determine which members of the exemplar base are most likely to affect the gender assignment of pan, and also to calculate the extent of analogical influence exerted.

---

2This example is a simplification; see (Eddington, 2002b) for a complete AM treatment of this phenomenon.
This is accomplished by calculating heterogeneity. Heterogeneity is determined by comparing the number of disagreements in the supracontextual and subcontextual analyses. If there are more disagreements in the supracontextual analysis, the supracontext is heterogenous, and its members are eliminated from consideration as possible analogs. If the number of disagreements does not increase, the supracontext is homogenous.

Words belonging to homogenous supracontexts comprise the analogical set (Table 4). In the example under consideration, disagreements increase in the supracontexts –––, and – a –. Therefore, their members are eliminated from consideration. Cal, and rey appear exclusively in these heterogenous supracontexts. As a result, they do not form part of the analogical set. The word plan is also a member of both –––, and – a –, however, it is also a member of the homogenous supracontexts – a n, and – – n, so it will still be available to influence pan.

It should not be surprising that rey would be eliminated through heterogeneity; it has no phonemes in common with pan. However, consider the words ron and cal. Both share only one feature with pan, yet heterogeneity eliminates only cal, and not ron. This is due to the fact that ron appears in the supracontext – – n, and all of the members of that supracontext are masculine, therefore, there is no disagreement. Cal F, on the other hand, competes with masculine words in all of the supracontexts in which it appears.

The analogical set contains all of the exemplar items that can possibly influence the gender assignment of pan. There are two methods for calculating the influence of the analogical set on the behavior of the given context. One is to assign the most frequently occurring behavior to the given context (selection by plurality). Of the 18 pointers in the set, 14 point to masculine. Therefore, selection by plurality would assign masculine gender to pan. The other method (random selection) involves randomly selecting a pointer, and assigning the behavior of the word indicated by the pointer to the given context. In this case, the probability of masculine gender assignment would be 77.78% (14/18).

### 3 Situating AM

The AM theoretical construct has been implemented computationally through several generations of application programs. Pascal code for the first version was listed as an appendix in (Skousen, 1989). A subsequent version was implemented in C. In recent years the system was reimplemented as a Perl module; extensive details on how to use the current version are available in (Skousen et al., 2002).

The system assumes *a priori* labeled input instances, and thus only works in supervised learning mode. Each exemplar must be represented as a fixed-length vector comprised of features that may or may not eventually be relevant to the phenomenon being modeled. The features themselves may be of fixed or variable length, as long as they are appropriately delimited. An outcome is specified for each exemplar instance vector. Vector features are purely symbolic and nominal, not admitting continuous values. When the use of continuous-valued features is needed (e.g. integers or real numbers),

<table>
<thead>
<tr>
<th>Supracontext</th>
<th>Subcontexts in Supracontext</th>
<th># Subcontextual Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>p a n</td>
<td>pan</td>
<td>0</td>
</tr>
<tr>
<td>p a –</td>
<td>pan, *pañ</td>
<td>2</td>
</tr>
<tr>
<td>p – n</td>
<td>pan, pàn</td>
<td>0</td>
</tr>
<tr>
<td>– a n</td>
<td>pan, pàn</td>
<td>0</td>
</tr>
<tr>
<td>p – –</td>
<td>pan, pàn, *pañ, *pañ, pàn</td>
<td>2</td>
</tr>
<tr>
<td>– a –</td>
<td>pan, pàn, *pañ, *pañ, pàn</td>
<td>2</td>
</tr>
<tr>
<td>– – n</td>
<td>pan, pàn, pàn, pàn, pàn, pàn</td>
<td>0</td>
</tr>
<tr>
<td>– – –</td>
<td>pan, pàn, pàn, *pañ, pàn, pàn, pàn, pàn</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Subcontextual analysis.
<table>
<thead>
<tr>
<th>Supracontexts</th>
<th>Words in Supracontext</th>
<th>Pointers</th>
<th># of Subcontextual Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>p a n</td>
<td>none</td>
<td>none</td>
<td>0</td>
</tr>
</tbody>
</table>
| p a –         | par M, paz F         | paz F > paz F  
par M > par M  
*paz F > par M  
*par M > paz F | 2                               |
| p – n         | none                 | none     | 0                               |
| – a n         | plan M               | plan M > plan M | 0                               |
| p – –         | par M, paz F         | paz F > paz F  
par M > par M  
*paz F > par M  
*par M > paz F | 2                               |
| – a –         | plan M, cal F, par M, paz F  
plan M > plan M  
*plan M > cal F  
plan M > par M  
*plan M > paz F  
*plan M > cal F  
*cal F > par M  
*cal F > paz F  
*cal F > cal F  
paz F > cal F  
paz F > plan M  
paz F > par M  
(cal F > paz F)  
(paz F > par M)  
*cal F > plan M  
*cal F > par M  
*cal F > paz F  
*cal F > cal F)  
paz F > cal F  
paz F > plan M  
paz F > par M  
(not all shown; all the rest involve agreement) | 6                               |
| – – n         | plan M, tren M, ron M | (not shown) | 0                               |
| – – –         | plan M, tren M, ron M, cal F, paz F, par M, rey M | (not shown) | 12                              |

Table 3: Supracontextual analysis.

the features must be symbolically quantized within pre-specified ranges in order to be meaningful to the system. More details on how to choose, quantize, and encode exemplar vectors is given in (Skousen et al., 2002).

As the system loads up the feature vectors, it builds the contextual lattice as sketched in the previous section. Once the lattice has been constructed, the system can process incoming test items to determine their outcome(s). Each test item consists of a feature vector containing the same number of features as the exemplar vectors do. Each test item also has the appropriate outcome specified, if known, or else some placeholder symbol. Each test item is matched against the lattice, its analogical set computed, and outcomes reported. AM, unlike some other machine learning approaches, can report all possible outcomes, and associates each with a probability rating or confidence rating reported as a percentage.

3.1 Sample implementations
AM has been applied to a number of different linguistic phenomena in several different languages. Although a comprehensive enumeration is not possible here, a representative sampling of applications gives an appreciation for the breadth and depth of coverage possible.

The earliest documented AM results with language data—and indeed the preponderance of work since—focus primarily on phonology, morphology,
Table 4: Analogical set non-empty homogenous supracontexts.

<table>
<thead>
<tr>
<th>Homogenous Supracontext</th>
<th>Words in Supracontext</th>
<th>Pointers</th>
<th># of Pointers to Masculine</th>
<th># of Pointers to Feminine</th>
</tr>
</thead>
<tbody>
<tr>
<td>pa –</td>
<td>par M, paz F</td>
<td>par M &gt; par M</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>– a n</td>
<td>plan M</td>
<td>plan M &gt; plan M</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>p – –</td>
<td>par M, paz F</td>
<td>par M &gt; par M</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>– – n</td>
<td>plan M, tren M, ron M</td>
<td>(not shown)</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

and lexical selection. Early applications (Skousen, 1989) included predicting Finnish past tense forms, describing allomorphic variation (e.g. the /an distinction in English), and predicting lexical selection based on sociolinguistic variables (e.g. Arabic forms for “friend”). Subsequent efforts (Skousen et al., 2002) addressed such phenomena as German pluralization, Dutch compound linkers, Spanish gender on nouns, stress in Dutch, and Turkish consonantal alternation.

Most of these implementations involve feature vectors that specify relatively low-level linguistic features such as sound segments, phonetic or orthographic environments, syllable structure, and word boundaries. Often the information for constructing these features comes from widely-used annotated language resources such as lexicons or corpora. For example, several English, German, and Dutch applications have used features partially derived the CELEX lexical database (Baayen et al., 1993); others have used WordNet (for English) (Fellbaum, 1998), TELL (for Turkish)³, and LEX-ESP (for Spanish) (Gallés et al., 2000).

With the advent of large-scale corpora, both annotated and unannotated, increasing amounts of information are available to language modelers. Recent AM efforts have benefited from speech and text corpora. The popular TIMIT speech database provided exemplars for recent work on flapping in English (Eddington, 2007). In probably the first AM application to diachronic language description, a recent paper (Chapman and Skousen, 2005) used the Helsinki corpus⁴ for modeling the historical development of English negative prefixes.

Increasingly, modelers are using their own experimentally obtained results to compare against more standard resources. The Arabic lexical selection example involved only features encoding novel researcher-collected situational observations. A recent paper on the comparative form of English adjectives (Elzinga, 2006) matches human judgments against occurrences from the internet collected via the Google search engine.

The choice of which features to include in the exemplar and test instances is an important issue for the AM user, as it is for all exemplar models. Since features have to be enumerated and encoded a priori, at least some of the features chosen should reflect salient properties of the phenomenon in question. This does not presuppose a prescient omniscience; it is not the case that all and only the relevant features must appear in the vectors. Any features that end up not contributing to the outcome simply constitute irrelevant overhead. Of course, in an exponential system such overhead should be avoided wherever possible.

3.2 Comparison to other approaches

The algorithm sketched in Section 2 is what sets the AM paradigm apart from other machine learn-
ing and language modeling approaches. The system does not employ sub-symbolic representations or continuous-valued features and thus differs in fundamental ways from connectionist systems. The computation of analogical (dis)similarity resembles in many ways nearest-neighbor approaches, though in some cases a “gang” of similar items, none of which is the nearest neighbor, may conspire to overcome the nearest neighbor and impose its own outcome. These gang effects are difficult to identify, though some have been discussed in the literature, and represent a unique aspect of AM results. When gang effects are not present, AM processing generally obtains results comparable to memory-based learning (e.g. in TiMBL). The latter, though, allows for continuous-valued features.

Another aspect where AM and some approaches differ is that AM considers each feature in any given vector of equal weight; only across the exemplar base do features emerge as more or less relevant. In this respect AM does not perform feature weighting in a separate processing stage across the exemplar base before testing new items. Similarly, no dependencies can be specified explicitly across features.

Extensive comparisons have been made between AM and other machine learning approaches including memory-based learning (Daelemans et al., 2004), version spaces, neural networks, and Optimality Theory; see (Skousen et al., 2002) for complete details.

The concept of an analogical set with quantifiable support is also unique to AM, and its relevance is being explored. For example, recent work (Chandler, 2005) has shown how AM results can map onto response times in certain types of psycholinguistic experiments. Eddington (2007 forthcoming) demonstrates how AM is able to mirror the results of psycholinguistic production studies that are presented elsewhere as evidence against analogy in processing regular morphology.

In fact, another notable feature of AM is its ability to model human language errors and variants that also result from analogy. Skousen, in early AM work, discusses how “leakage” (i.e. erroneous AM guesses) often tends to follow observable usage errors (Skousen, 1989).

Eddington’s (2002a) simulations of gender assignment in Spanish yielded similar errors rates to those of children of different ages, and in a simulation of Spanish diminutives (Eddington, 2002a), 52% of the incorrect predictions made by AM were actually found to be attested alternative diminutive forms.

A comparison of corpus data for Russian verbs of motion—and subsequent modeling in AM combining lexical, morphological, and semantic variables—showed a tendency for the system’s errors to coincide with those of language learners (Dodge and Lonsdale, 2006).

Certainly further comparison of the behavior observed in AM modeling tasks versus human performance constitutes a promising future direction to pursue.

When considering the place of AM in the machine learning realm, one area deserves considerable more effort. The natural language processing (NLP) field has profited greatly from machine learning, in part due to the increased availability of large-scale language resources. Though early work in AM illustrated the promise of the approach to typical NLP tasks (Jones, 1996), the paradigm has been underrepresented in recent NLP shared tasks and competitive evaluations addressing such problems as word-sense disambiguation, part-of-speech tagging, shallow parsing, semantic role labeling, and phoneme classification. Thanks to recent developments in the implementation as discussed in the next section, we expect this situation to improve in the future.

4 Recent progress in AM

The AM system has undergone substantial development since the last documented release of the system (Skousen et al., 2002), and this has permitted applications in areas that would have been previously impossible. In this section we survey some of the recent AM developments and consequent research they have enabled.

4.1 Algorithmic refinement

As explained above, the AM system’s basic data structure is a fully connected lattice representing (dis)agreements among data items. The original algorithm for handling the lattice involved a straightforward but exhaustive traversal and was hence very costly. The most current version of the system in-
volves a completely new, but equivalent, method of extracting information from the lattice.

The crucial insight is to focus on heterogeneity instead of homogeneity. For each supracontext in the lattice, heterogeneity can be locally determined by looking for plurality of outcomes and plurality of subcontexts. When heterogeneity is detected, all more general supracontexts can be automatically discarded from that point on.

Implementing this insight has resulted in a substantial speed-up in processing time, and the ability to handle a greater number of features, exemplars, and test items. Figure 1 illustrates the run time for different versions of the AM program on a typical task (English part-of-speech tagging) involving 50,000 exemplars, 1,000 test items, and 58 outcomes. It shows that the previous version of the system grew exponentially as the number of variables grew (though the maximum number of variables possible was fixed at 28); the algorithm also performed particularly poorly for a vector size of 3 features. This contrasts with the current version of the system, which runs largely linearly (on a dual-processor Sun Sparc Blade 25000) until the current maximum vector size of 59 features, at which point the exponential curve begins. A new 64-bit version of the system continues to behave linearly even at 59 features.

The rewrite of the algorithm also incorporated more decomposition in the handling of the lattice, in theory laying the groundwork for a truly parallel implementation. To date, though, no concurrent parallel implementation has been developed.

Even with this improvement, exponentiality still poses a problem, if only theoretically. Two independent efforts have been pursued recently to address the processing bottleneck issue. One involves using Monte Carlo sampling to estimate the results that traditional AM processing provides (Johnsen and Johansson, 2005). The other method is to compute the dissimilarities via a quantum algorithm (Skousen, 2005). To discuss either would be beyond the scope of this paper.

4.2 Modularization

The newest version of the AM implementation has been repackaged as a Perl module. This was done primarily to improve flexibility in using the system. Its deployment within Perl supports better cross-architecture functionality. The provision of a number of processing hooks allows for various user-programmable add-on procedures for customizing
the input, matching, and output stages of processing.

One example of the use of customized output involves the integration of AM with a cognitive modeling system for natural language called NL-Soar. In this work, which involves parsing, the cognitive agent must occasionally resolve prepositional-phrase attachment ambiguities (Lonsdale and Manookin, 2004; Manookin and Lonsdale, 2004); it does so by invoking AM to resolve the ambiguity using a PP-attachment exemplar set previously mined by others⁵. This integration would have been quite difficult with prior versions of the AM system.

Some modeling problems are sequential in nature, meaning that a particular test’s outcome might be dependent on that of a preceding test's outcome: dynamically chaining results from one test to another is necessary in such cases and now possible with AM. Similarly, collecting results of individual tests, integrating them, and reporting them in another format is also now possible in the current AM system.

For example, recent work on Romanizing Farsi names (Lonsdale, 2006) profited from the improved AM system input/output hooks. Input to the system consisted of an exemplar base of several thousand Farsi names in Arabic script form⁶, and their Romanized counterparts; often there is more than one way to Romanize such names. The system processes each name, for which all possible Romanizations are generated from successive analogical sets. In addition, a client-specified metric measuring the confidence of each possible outcome can also be generated via the output hook and further post-processing. Figure 2 illustrates sample results. The first input, bhnAm, represents the Arabic-script name; two possible Romanized forms (Bahnaam and Behnaam) are posited by the system, with respective confidence scores of 436.044 and 402.424. Note that the third input name has ellipsis dots (“…” ) representing the Divine Name, which is pronounced but not written; even in this case the system can compute the proper Romanized substitutions via analogy.

---


⁶For convenience here we do not use actual Arabic-script characters for the input but rather transliterations.

Figure 2: Sample summary report from Farsi name Romanization process showing inputs and outputs with alternatives scored per custom algorithm.

4.3 Lattice visualization

For many users of the AM system, insight into the operation of the AM algorithm has proven difficult, with the system operating as a black box. A new visualization tool allows users to load up their exemplars, enter test items, and step through the lattice incrementally to view the effects of matches, homogeneity/heterogeneity, and gangs.

Figure 3 shows that testing the item pan against the exemplars in the supracontext p ā n results in non-determinism (specified by par M and paz F) in a homogeneous supracontext.

4.4 Architectural improvements

Loading a large number of exemplar items is time-consuming. This is keenly felt in sequential implementations, where the outcome of one test might depend on the results of a previous one (cf. the Farsi example described above).

To alleviate this problem, a new implementation provides a client/server infrastructure. The exemplar data can be loaded on a server machine once, and the server stands by for test item queries (over a socket)
from client machines.

4.5 User tools

Recent use of the system by users has resulted in two other tools worthy of note.

Development of exemplar and test item feature vectors can sometime result in ill-formed vectors which are difficult to locate. A new Perl script allows users to check their feature files and receive a report of which vectors do not comply to formatting specifications.

For users unable to configure the basic system parameters within the Perl code, use of the system can be awkward. A basic Perl/TK graphical user interface has been developed to provide users top-level control to the overall system, allowing them to select pertinent parameters and input files. In this way the users will not need to modify any Perl code.

5 Conclusions

The AM algorithm has been applied to a number of different language phenomena where small numbers of variables and long processing times were not an issue. For psycholinguists, the most important factors in a model are often its ability to mirror the responses and errors of language speakers, along with the psychological plausibility of the model. In NLP, on the other hand, processing speed and the ability to incorporate large numbers of variables are often more highly favored. Given the advances in the current implementations of the AM algorithm we trust that AM will become a more attractive tool for a wider range of linguistic applications.

Acknowledgements

We would like to thank Theron Stanford for his programming support work and comments.

References


From Exemplar Theory to Population Coding and Back.
An Ideal Observer Approach.

Laurent Bonnasse-Gahot†,∗ and Jean-Pierre Nadal‡,‡
†Centre d’Analyse et de Mathématique Sociales
UMR 8557 CNRS-EHESS
Ecole des Hautes Etudes en Sciences Sociales
54 bd. Raspail, F-75270 Paris Cedex 06
‡Laboratoire de Physique Statistique
UMR 8550 CNRS-ENS-Paris 6-Paris 7
Ecole Normale Supérieure
24 rue Lhomond, F-75231 Paris Cedex 05
∗Corresponding author. Email: lbg@ehess.fr

Abstract
Exploiting the analogy between exemplar models and population coding schemes, we characterize, by means of information theoretic tools, the efficiency of a large but finite population of cells coding for a discrete set of categories (e.g. vowels). The optimal code is shown to typically allocate more cells to class boundaries than to regions further apart. We then discuss the predicted perceptual consequences and review existing exemplar models in the light of our general results.

Keywords: exemplar models, population coding, speech perception, information theory, psychophysics.

1 Introduction
Exemplar models (Hintzman, 1986; Nosofsky, 1986) originally stem from the field of psychology as general models of perception and categorization. They have subsequently been applied to speech perception (Lacerda, 1995; Johnson, 1997) and extended to speech production (Pierrehumbert, 2001, 2003). Although a neuroscientific interpretation has sometimes been mentioned (Lacerda, 1998), such an approach has never been seriously exploited. For instance Kruschke’s model, "motivated by a molar-level psychological theory" (Kruschke, 1992), despite a terminology partly borrowed from neuroscience – e.g. activation, receptive field –, remains in line with traditional connectionism, as suggested by the use of node instead of neuron or cell.

In this paper, within the framework of speech perception, we propose to take seriously the hypothesis that exemplar models can be given a direct interpretation in term of neural representation. Taking advantage of a recent literature in neuroscience, and making use of standard tools from information theory (see, e.g., Blahut, 1987; Cover and Thomas, 2006), we show not only that this neuroscientific approach is plausible but also that it makes it possible to study in a very general way the optimal neural coding of categories (e.g. vowels), independently of any assumptions about learning or decoding method. This, in turn, will be shown to have consequences for exemplar theory.

This paper is organized as follows. Section 2 sums up the main mathematical result derived in Bonnasse-Gahot and Nadal (2007). In subsection 2.1, we first give a description of the model we use, pointing out its links with both exemplar theory and theoretical neuroscience. We then exhibit in subsection 2.2 the main formula as well as the following predictions. In section 3 we present the perceptual interpretation and consequences of our result. Section 4 considers existing models in the light of our results, and the last section finally gives the concluding remarks.

2 Population coding of categories
2.1 Model description
We assume given a discrete set of categories \( \mu = 1, \ldots, M \) (such as phonemes, but our results are applicable to other modalities than speech as well), with probabilities of occurrences \( q_\mu \geq 0 \),
so that $\sum_\mu q_\mu = 1$. Each category defines a density distribution $P(x|\mu)$ over the continuous stimulus space (see Pierrehumbert, 2003, p. 119). Along with classical exemplar views, the stimulus space is assumed to be of finite dimensions, those dimensions being those relevant to speech perception (e.g. in the case of vowels, the perceptual dimensions might be the fundamental frequency F0 and the first formants F1, F2, F3). For the sake of clarity, we consider here a one dimensional case: $x \in \mathbb{R}$ (the general case of $x \in \mathbb{R}^K$ is presented in Bonnasse-Gahot and Nadal, 2007).

Playing the role of $N$ stored exemplars or $N$ (Kruschke’s) hidden nodes, we consider a population of $N$ neurons. Each neuron $i$ has an activity specific to a location within the input space, with a mean response given by its tuning curve $f_i(x)$, centered around a value $x_i$ (the preferred stimulus of cell $i$, which can be considered as the stored exemplar), and with a width $a_i$. A typical tuning curve is given by a bell-shaped function, such as

$$f_i(x) = F_i \exp\left(-\frac{(x - x_i)^2}{2a_i^2}\right)$$

In a standard exemplar model, one would have a uniform value $a_i = a$ of the width. Here, the heterogeneity in the widths $a_i$ allows for local deformations of the perceptual space defined by the output of the neuronal population (we will discuss this point in section 3).

We assume that the responses of the neurons (given $x$) are not correlated, so that the overall activity $r = \{r_1, \ldots, r_N\}$ has a factorized probability density function:

$$P(r|x) = \prod_{i=1}^N P_i(r_i|x)$$

with thus

$$\sum_{r_i} P_i(r_i|x)r_i = f_i(x).$$

For the numerical illustration of our results, we will consider that, given an input $x$, the activity (number of spikes) $r_i$ of the $i$th neuron is generated according to a Poisson statistics with mean rate $f_i(x)$, that is:

$$P_i(r_i|x) = \frac{(f_i(x))^{r_i}}{r_i!} e^{-f_i(x)} \quad (2.1)$$

This Poisson model is taken here for both its mathematical simplicity and its biological plausibility (see e.g., Tolhurst et al., 1983; Softky and Koch, 1993).

Such a coding is a typical instance of population coding (see e.g., Pouget et al., 2000), a strategy widely used in the brain that consists in encoding information by large assemblies of neurons. Two well-known examples are given by the representation of movement direction in the primate motor cortex (Georgopoulos et al., 1986), or by the head-direction cells in rats (Taube et al., 1990). A particularly relevant example here is the inferotemporal cortex in the monkey, which has been shown to be a site for object recognition (see Tanaka, 1996, for a review) and classification. There, population coding is a strategy widely used (e.g. Young and Yamane, 1992; Vogels, 1999), and has already be given an exemplar-based interpretation (Logothetis et al., 1995; Sigala and Logothetis, 2002; Sigala, 2004).

We are interested in quantifying the coding efficiency of such a (neural) representation, and in characterizing the optimal one. Optimality is defined as minimizing the probability of error of an ideal observer during a task of classification. We do not address the question of learning or decoding. We thus do not assume any particular type of learning process nor any decoding method, so that our results remain general and can be applied to any model that shares the same basic assumptions (e.g. Nosofsky, 1986; Kruschke, 1992; Lacerda, 1995; Johnson, 1997).

2.2 Results

**Mutual information.** In theoretical neuroscience, population coding efficiency has been computed by means of information theoretic tools, in the case of both continuous and discrete stimuli (Seung and Sompolinsky, 1993; Brunel and Nadal, 1998; Kang and Sompolinsky, 2001). We perform a similar analysis in the present context of categorical perception.

A relevant quantity is the mutual information that measures the statistical dependency between two variables. Here, we want to compute the mutual information between the set of categories and the neural representation. Maximization of this quantity (which can be the result of learning or adaptation) will have the consequence of minimizing the probability of misclassifying an incoming
The mutual information between the categories \( \mu \) and the neural activities \( r \) is defined by (Blahut, 1987):

\[
I(\mu, r) = \sum_{\mu=1}^{M} q_\mu \int d^N r \ P(r|\mu) \log \frac{P(r|\mu)}{P(r)}
\]

(2.2)

where \( P(r) \) is the probability density function (p.d.f.) of \( r \):

\[
P(r) = \sum_{\mu=1}^{M} q_\mu P(r|\mu).
\]

(2.3)

This quantity \( I(\mu, r) \) is positive and, by virtue of the data-processing theorem (see e.g. Blahut, 1987), it is upper bounded by the information \( I(\mu, x) \) conveyed by the sensory input \( x \) about \( \mu \). Under mild assumptions one can show that

\[
\lim_{N \to \infty} I(\mu, r) = I(\mu, x).
\]

**Large but finite population.** In Bonnasse-Gahot and Nadal (2007), we show that for finite but large \( N \) the leading correction to this \( N \to \infty \) limit is given by:

\[
I(\mu, x) - I(\mu, r) = \frac{1}{2} \int dx \ p(x) \ F_{\text{code}}(x) F_{\text{cat}}(x)
\]

(2.4)

where \( p(x) \) is the p.d.f. of the stimulus \( x \): \( p(x) = \sum_\mu q_\mu P(x|\mu) \), and \( F_{\text{code}}(x) \geq 0 \) is the Fisher information characterizing the sensibility of \( r \) with respect to small variations of \( x \) (see, e.g., Blahut, 1987):

\[
F_{\text{code}}(x) = -\int d^N r \ P(r|x) \frac{\partial^2 \ln P(r|x)}{\partial x^2}
\]

(2.5)

and \( F_{\text{cat}}(x) \geq 0 \) is the Fisher information characterizing the sensibility of \( \mu \) with respect to small variations of \( x \):

\[
F_{\text{cat}}(x) = -\sum_{\mu=1}^{M} P(\mu|x) \frac{\partial^2 \ln P(\mu|x)}{\partial x^2}
\]

(2.6)

which can also be written as

\[
F_{\text{cat}}(x) = \sum_{\mu=1}^{M} \frac{P'(\mu|x)^2}{P(\mu|x)}
\]

(2.7)

In (2.7), \( P'(\mu|x) \) denotes the derivative of \( P(\mu|x) \) with respect to \( x \).

The Fisher information \( F_{\text{code}}(x) \) is specific to the coding stage \( x \to r \): it tells how well the neural code can discriminate nearby sensory inputs. The term \( F_{\text{cat}}(x) = \sum_{\mu=1}^{M} P'(\mu|x)^2/P(\mu|x) \) is specific of the sensory encoding \( \mu \to x \) and thus does not depend on the neural code: it tells how the statistics in the input space are well correlated or not to the categories.

**Main qualitative consequences.** Typically, an identification function \( P(\mu|x) \) has an S-shape, whose slope \( |P'(\mu|x)| \) is largest near the boundaries between categories. This entails that the quantity \( F_{\text{cat}}(x) = \sum_{\mu=1}^{M} P'(\mu|x)^2/P(\mu|x) \) is greater in these regions. If the code is to be optimized, we therefore expect, as the number \( N \) of neurons is limited, Fisher information \( F_{\text{code}}(x) \) to be greater between categories than within (see Eq. 2.4). As a consequence, more cells will be devoted to these regions of overlap compared to regions where only one category dominates. This is reminiscent of the Support Vector Machine approach (Cortes and Vapnik, 1995), a technique very popular in machine learning, that identifies exemplars closest to the class boundary as being the most crucial ones for a given classification task. Besides, this result seem to find support in functional imagery and neuro-physiology. First, using functional imagery methods, Guenther et al. (2004) show both for speech and non-speech sounds that category learning entails neural activity in the auditory cortex to be higher in response to stimuli close to the boundary (‘non-prototypical’ stimuli) than in response to prototypical stimuli (that lie in a more central region of a given category) (see also Guenther and Bohland, 2002, and section 4). Second, in the neuro-physiology of the inferotemporal cortex of the monkey brain, Freedman and colleagues found that category learning leads to a distribution of preferred stimuli mainly peaked at the class boundary: almost half of all the recorded neurons have indeed their preferred stimuli located at the class boundary (Knoblich et al., 2002, Fig. 12).

**Numerical Illustration.** Figure 1 shows a numerical example involving two categories, whose distributions are shown in Fig. 1.A, and \( N = 15 \) neurons initially equidistributed. The optimal \( \{x_i\}_{i=1}^{N} \) and \( \{a_i\}_{i=1}^{N} \) are obtained by numerically
maximizing (using simulated annealing) the difference \( I(\mu, r) - I(\mu, x) \) given by equation 2.4. Fig. 1.B shows the resulting tuning curves. As expected, Fisher information \( F_{\text{code}}(x) \) (plotted in Fig. 1.C) is the greatest at the boundary between the two categories, and more neurons are allocated in this region compared to regions further apart. Note also that the tuning curves are sharper near the boundary, \( i.e \) the width \( a_i \) of the corresponding cells is narrower than the width of cells away from the boundary. One could see that this width plays the role of the inverse of the ‘attentional weight’ found in classical exemplar-based models (Nosofsky, 1986; Kruschke, 1992). In other words, more local ‘attention’ is devoted to the class boundary, further sensitizing the neuronal population to this region. Note that this is a collective effect, coming from having both an heterogeneous set of widths (each cell \( i \) having its own \( a_i \)) and a specific distribution of preferred stimuli.

3 Towards an explanation of categorical perception

As previously stated, if the code is optimized, the Fisher information \( F_{\text{code}}(x) \) is greater at the boundary between categories than within (Eq. 2.4). The Fisher information \( F_{\text{code}}(x) \) is linked to the discriminability \( d' \) (from signal detection theory, and commonly used in psychophysics; see e.g. Green and Swets, 1988) of two stimuli \( x \) and \( x + \delta x \) according to (Seung and Sompolinsky, 1993):

\[
d' = |\delta x| \sqrt{F_{\text{code}}(x)}
\] (3.9)

Thus, a perceptual consequence of maximizing mutual information between neural responses and categories, under the constraint of a fixed number of cells, is that discriminability \( d' \) will be greater at the boundary between categories, a phenomenon traditionally called categorical perception (Harnad, 1987).

Categorical perception was first described within the field of speech perception as an innate process specific to human speech that implied high discriminability between items from different (phonemic) categories and zero discriminability within a category (Liberman et al., 1957). This strong version of categorical perception was subsequently undermined: not only this phenomenon can be acquired (Abramson and Lisker, 1970; Francis and Nusbaum, 2002) but it is also not specific to speech (Goldstone, 1994; Livingston et al., 1998; Özgen and Davies, 2002) nor to human (Kuhl and Padden, 1983; Kluender et al., 1998). Moreover, the all-or-none effect on discriminability was never found experimentally: within-category differences are discriminable. Our result fits well into this framework, for it can apply to any modality, might be induced by learning, and does not assume anything specific to human. Besides, the discriminability within a category is not zero.

Figure 1: One-dimensional example involving two Gaussian categories. (A) Probability distributions of the two categories. (B) Optimal tuning curves. (C) Difference between the Fisher information \( F_{\text{opt}}^{\text{code}} \) for the optimal code, and the Fisher information \( F_{\text{init}}^{\text{code}} \) for a equidistributed distribution of preferred stimuli.
Another way to present the effects induced by the adaptation of the neural configuration is in terms of compression/expansion of the perceptual space defined by the output of the neuronal population. If discriminability is higher (respectively lower) after learning than before, then the perceptual space can be seen as expanded (resp. contracted). Category learning might imply different outcomes. For example, using visual stimuli, Goldstone (1994) found *acquired distinctiveness* at the class boundary (increased between-categories differences), whereas Livingston et al. (1998) found *acquired similarity* (increased within-category similarity). In the case of the learning of new phonetic categories, Francis and Nusbaum (2002) reported both compression and expansion of the perceptual space. Whether category learning induce within-category compression and/or between-category expansion might depend on the initial ability of the neuronal population. In the numerical illustration given by figure 1, we see that, compared to the equidistributed initial configuration of preferred stimuli, there is *acquired distinctiveness* at the boundary and *acquired similarity* within categories. Such a warping of the perceptual space is related to a phenomenon found in speech perception literature called the *perceptual magnet effect* (Kuhl, 1991), stating that discriminability is lower around prototypical stimuli than around non-prototypical ones, even if the corresponding stimuli belong to the same category. These prototypicality effects (as well as frequency effects) will be more extensively studied in a forthcoming paper (Bonnasse-Gahot and Nadal, in preparation).

An important aspect of our model is that we do not need category labels so as to find categorical effects, which might shed light on one of the most disputed issues on categorical perception. This question, that finds its roots in the Whorffian hypothesis stating that our language shapes our vision of the world (Kay and Kempton, 1984), paradoxically concerns the very basis of this phenomenon: is categorical perception really perceptual? Views are divided. Some argue that this phenomenon is not perceptual after all but only results from the use of verbal labels (Roberson and Davidoff, 2000), whereas others maintain that categorization does alter perception (Goldstone et al., 2001; Notman et al., 2005). Our view follows the latter. Our result indeed gives an optimal bound on discriminability, based on a purely sensory level, and thus gives credit to a perceptual account for the two phenomena described above, namely categorical perception and perceptual magnet effect. Note, however, that we do not claim that other processes, such as top-down influences, memory effects, or labeling, might not intervene in discriminability judgments.

To conclude, our result indicates that categorical perception is not a mere by-product of category learning but serves a function, that is to minimize classification errors. This gives a quantitative theoretical support to the Native Language Neural Commitment posited by Kuhl (2004) stating that language experience induces neuronal modifications that aim at enhancing the features relevant for native language but entail difficulties in the learning of a foreign language (see, e.g., Kuhl et al., 1992; Iverson et al., 2003).

4 Comparison with existing models

In this section we want to reconsider existing models in the light of our results. Two main kinds of models, designed to account for categorical perception and/or perceptual magnet effect, are concerned: exemplar-based models (Lacerda, 1998; Goldstone et al., 1996) and neural maps (Bauer et al., 1996; Guenther and Gjaja, 1996; Guenther and Bohland, 2002). Note that all these models share the same basic assumptions concerning the architecture: a perceptual map is covered by ‘cells’ or ‘exemplars’ centered around a preferred stimuli with the responsiveness of their receptive field decreasing as the incoming stimulus moves away from the preferred stimulus. The models differ by how this map is decoded (e.g. with a specific additional layer), and/or by the learning algorithm used to build the map. In our case, we have characterized the coding efficiency of a map (see equation 2.4), independently of any learning process or decoding method. Taking as a reference the expected properties of an optimal code, we can thus compare our predictions with the empirical results obtained using specific algorithms.

As we have seen in section 2.2, a direct consequence of equation 2.4 is that category centers are represented by fewer cells (neurons,
exemplars) than category boundaries, which in turn explain why perceptual phenomena such as categorical perception or the perceptual magnet effect might arise (section 3). This is in line with results obtained by Goldstone et al. (1996), Bauer et al. (1996), Guenther and Bohland (2002) but in disagreement with models proposed by Guenther and Gjaja (1996) or Lacerda (1995, 1998). Let us review these two latter models in more depth.

Interestingly, Guenther and Gjaja (1996) model is the only one from the above list for which the learning algorithm is not specifically meant to solve a categorization task. Following traditional self-organizing map approach, the learning mechanism leads to a situation where the distribution of preferred stimuli of the coding cells follows the distribution of stimuli. This actually corresponds to a situation of density estimation (representing the distribution of $x$, instead of coding for the categories), and should lead to higher discriminability where the distribution of stimuli peaks, i.e. at the center of the categories, contrary to a situation of categorical perception. Guenther et al. (1999) have experimentally shown that the same distribution of stimuli can lead to opposite perceptual outcomes, depending on the training task (discrimination vs categorization). Guenther therefore proposed a different version so as to take into account the need for categorization. This time, the subsequent model allocates more cells to the boundary (Guenther and Bohland, 2002; Guenther et al., 2004), in agreement with our qualitative predictions.

Let us now turn to Lacerda’s model (Lacerda, 1995, 1998). It basically assumes that every encountered exemplar is stored, leading again to more cells around the modes of the distribution of stimuli, contrary to our result. In order to explain the perceptual phenomena discussed above, a discrimination measure involving the category label of the exemplars is introduced. This runs counter to the fact that labels might not be used in a discrimination task, especially in the case of the magnet effect for which it is assumed that all items belong to the same category. For instance, in Iverson and Kuhl (1995), subjects were conditioned to view all items as belonging to the same category, which might prevent them from using category labels, hence working on a more perceptual basis.

More generally, our result goes counter to the most common working hypothesis in computational exemplar models that all encountered exemplars are stored, leading to a higher density of cells at the center of a category. It also sheds light on the “head-filling up problem” (Kruschke, 1992; Johnson, 1997; Pierrehumbert, 2001) that the more traditional approach has to face, as the memory is of finite size. It indeed shows that some exemplars are more informative about categories than others, and that much compression can thereby be achieved at the center of a category.

5 Conclusion and Future Works

To sum-up, we first gave a neural interpretation of exemplar models in terms of population coding of categories, which makes it possible to address the question of optimal coding with information theoretic tools that have been shown to be relevant in computational neuroscience. We find that the mutual information between the activity of a neuronal population and a set of discrete categories is simply given by the average over the input space of the ratio of the Fisher information of the categories over the Fisher information of the neuronal population. As seen in section 2.2, the category related Fisher information is typically the greatest at the boundary between categories. As neural resources are limited, this entails that, if there is adaptation, the Fisher information of the neuronal population should also be the greatest between categories in order to minimize the probability of misclassifying an incoming stimulus. As this Fisher information is directly related to the discriminability, category learning implies better cross-category discrimination than within-category discrimination, which gives an explanation of categorical perception (see section 3).

As we have seen, we make no assumption on the learning mechanism nor on the decoding method. This makes it possible to evaluate existing models (section 4) and more generally to establish a groundwork for future work. A possible direction for future research will now consist in studying these mechanisms in more depth.

Let us first get into the question of decoding. Several methods such as population vector or maximum likelihood have already been proposed
in theoretical neuroscience, so that this point might not be a technical issue. More interestingly, one can ask whether decoding is necessary. Empirical research has indeed shown that categorical information as well as fine phonetic details within-categories are used by the listener (Miller, 1994; McMurray et al., 2002). In that spirit, our results might give clues on how two different levels of representation, one continuous and concrete (stimulus) and one discrete and abstract (categories), can be combined within the same neural representation. On the one hand, thanks to population coding, information about the stimulus and its detailed properties might be retrieved, but on the other hand, because of the neural modifications induced by category learning (language experience), this neural representation also contains information about the category the stimulus belongs to.

Concerning the question of learning, cross-language studies have demonstrated that infants of 6 months of age are already tuned to their native language, as in the case of the perception of vowels (Kuhl et al., 1992), well before they can talk or have acquired a lexicon. Moreover, many experiments have recently shown that both adult, infant, and animal listeners are able to extract distributional information from the input signal (Saffran et al., 1996; Guenther et al., 1999; Lotto, 2000). Maye et al. (2002) showed not only that 6 and 8-month old infants are sensitive to the statistical distribution of the speech sounds they hear in their environment but also that this in turn influences their discrimination ability. These results call for the study of unsupervised learning that would aim at attaining the ideal state described by our result. Such a study will have to face a paradoxical issue given by our result. More resources have indeed to be allocated to category boundaries that are typically regions where exemplars are rarely encountered. As a consequence, one might ask how information about the distribution of stimuli might be extracted from regions of low-frequency ‘traffic’. Beyond technical issues, this question is particularly interesting from a developmental perspective.

Acknowledgments

This work is part of a project “Acqlang” supported by the French National Research Agency (ANR-05-BLAN-0065-01). LBG acknowledges a fellowship from the DGA. JPN is a CNRS member.

We thank Sharon Peperkamp and Janet Pierre-humbert for introducing us to this topic and for valuable discussions. LBG is also grateful to Emmanuel Dupoux for numerous and stimulating discussions. We also thank three anonymous referees for useful comments and suggestions.

References


Liberman, A., Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech...


