

# Implicit schemata and categories in memory-based language processing

Antal van den Bosch (Antal.vdnBosch@uvt.nl)

Tilburg center for Cognition and Communication, Tilburg University  
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Walter Daelemans (walter.daelemans@ua.ac.be)

Computational Linguistics & Psycholinguistics Research Centre, University of Antwerp  
Prinsstraat 13 (L), 2000 Antwerpen, Belgium

## Abstract

Memory-based language processing (MBLP) is an approach to language processing based on exemplar storage during learning and analogical reasoning during processing. From a cognitive perspective, the approach is attractive because it does not make any assumptions about the way abstractions are shaped, and does not make any a priori distinction between regular and exceptional exemplars, allowing it to explain fluidity of linguistic categories, and irregularization as well as regularization in processing. Schema-like behavior and the emergence of categories can be explained in MBLP as by-products of analogical reasoning over exemplars in memory. By reviewing a number of cases in morpho-phonology and then zooming in on memory-based language modeling, operating beyond the word level, we show how abstractions arise implicitly in a memory-based framework. We critically discuss the differences between the MBLP approach and other frameworks that do assume some systemic form of abstraction (e.g. prominence hierarchies in syntactic tree fragments).

**Keywords:** memory-based language processing; generalization; abstraction

## Memory-Based Language Processing

Memory-based language processing, MBLP, is based on the idea that learning and processing are two sides of the same coin. Learning is the storage of examples in memory, and processing is similarity-based reasoning with these stored examples. Although we have developed a specific operationalization of these ideas (Daelemans & Van den Bosch, 2005), they have been around for a long time. We first provide an overview of similar ideas in computational and cognitive linguistics.

### MBLP and computational linguistics

We see MBLP as the implementation of the example-based strand of linguistic theories developed throughout the twentieth century, from Saussurean analogical reasoning to example-based models of human language processing.

MBLP finds its computational basis in the classic  $k$ -nearest neighbor classifier (Cover & Hart, 1967). With  $k = 1$ , the classifier searches for the single example in memory that is most similar to  $B$ , say  $A$ , and then copies its memorized mapping  $A'$  to  $B'$  (as visualized schematically in Figure 1). With  $k$  set to higher values, the  $k$  nearest neighbors to  $B$  are retrieved, and some voting procedure (such as majority voting) determines which value is copied to  $B'$ .

Here we already reach a crucial limit of state-of-the-art machine learning and probabilistic methods, and also of

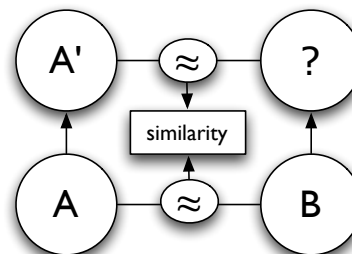


Figure 1: Saussurian analogy. Given the proportional analogy  $A:A'::B:B'$ , with  $B'$  missing, we need to find a  $B'$  that fits the analogy.

memory-based learning; in most general cases in natural language processing it is impossible to generalize to entire  $B'$  sequences due to the immense sparseness of subsequences in language longer than three or four words. Hence, except for a particular branch of work on full analogical proportions on which we expand in the *Related approaches* section, a full Saussurean analogical reasoning process is nowadays typically broken into smaller subtasks where subsequences of  $B'$  are computed separately; subsequently, a global search may then be applied to the set of partial solutions to find the most likely complete  $B'$ .

In sum, a “subsequencing” version of Saussurean analogy is the basic model for the majority of present-day mainstream natural language processing algorithms, and memory-based learning is one of the few models that base their decisions on extrapolations from single examples, or  $k$  examples, rather than on amalgamate models that abstract away from single examples in principle.

### MBLP and cognitive linguistics

Exemplar-based models have been proposed in psychology, more specifically in studies of human categorization, and have been argued to produce a generally good fit of human behavior and errors (E. Smith & Medin, 1981; Nosofsky, 1986; Estes, 1994). These models assume that people represent categories by storing individual exemplars in memory rather than rules, prototypes, or probabilities. Categorization decisions are then based on the similarity of stimuli to these stored exemplars. Evidence for the psychological relevance of exemplar-based reasoning remains impressive. Recently, even the very assumption of *fixed*, perma-

ment categories (however represented) has come under fire by theories favoring a *dynamic construal* approach in which concept formation is claimed to be based on past and recent experiences represented in memory, combined with current input (L. Smith & Samuelson, 1997). This type of context-dependent, memory-based category emergence fits MBLP very well.

One recent approach to linguistics, *usage-based* models of language, represented by cognitive linguists such as Ronald Langacker, Joan Bybee, Adele Goldberg, William Croft and many others (Croft & Cruse, 2003; Goldberg, 2006), bases itself at least in part on the psychological categorization literature and on some of the pre-Chomskyan linguistic approaches discussed earlier. Some of the properties shared by the heterogeneous set of usage-based theories are reminiscent of the MBLP approach. Most importantly, the usage-based approach presupposes a bottom-up, maximalist, redundant approach in which patterns (schemas, generalizations) and instantiations are supposed to coexist, and the former are acquired from the latter. MBLP could be considered as a radical incarnation of this idea, in which *only* instantiations stored in memory are necessary. Other aspects of cognitive linguistics, such as the importance of frequency, and experience-based language acquisition (Tomasello, 2003) fit MBLP as well.

### Related approaches

There is a close relation between memory-based language processing and two schools of analogical (computational) modeling of language: those of Royal Skousen and colleagues (Skousen, 1989; Skousen, Lonsdale, & Parkinson, 2002), and a group that could be identified as the “French analogical proportionists”; Yves Lepage, François Yvon, and colleagues (Lepage & Shin-ichi, 1996; Yvon & Stroppa, 2007; Langlais, Yvon, & Zweigenbaum, 2009). The relation of memory-based language processing with Skousen’s analogical modeling work has been discussed earlier in Daelemans (2002). The approach has been applied mainly in the phonology and morphology domains. Empirical comparisons have never shown important accuracy or output differences between the two approaches (Eddington, 2002; Daelemans, 2002).

In the work of the French analogical proportionists, Lepage, Yvon, Langlais, Stroppa, and other colleagues stress the importance of adhering to the full proportional analogical reasoning that De Saussure proposed, where the sequences in the proportional relation  $A:B::A':B'$  are truly the full sequences in all their complexities (Lepage & Shin-ichi, 1996).

A somewhat more indirect family relationship exists between analogical methods on the one hand, and example-driven stochastic structural models on the other hand, such as DOP, data-oriented parsing (Bod, Scha, & Sima’an, 2003). Data-oriented models of the DOP type are essentially single probabilistic models, as they divide a global probability mass over a large population of labeled tree fragments. Yet, Scha, Bod, Sima’an and colleagues do stress in their work on

DOP the reliance of the method on individual examples. A DOP parsing operation can be traced back to the set of individual parsing tree fragments involved in the process. It has furthermore been observed in DOP models that removing individual examples on the basis of their rarity (their low frequency) hampers performance considerably (Bod, 1995), in line with our observations (Daelemans, Van den Bosch, & Zavrel, 1999).

If other differences such as probabilities versus “natural” frequencies and distances are ignored, a key difference between the DOP approach and the memory-based approach is the assumption in DOP that examples are fragments of hierarchical structures, while the standard version of the memory-based language processing model assumes no pre-defined structure. The DOP approach, and also the recent approach proposed by Post and Gildea (2009), implies a resolution mechanism in which found or activated fragments join and form a tree. It follows naturally that this approach is typically cast in the framework of a syntactic task. Despite its wide applicability from morpho-phonology to syntactosemantic processing, it is not the most straightforward solution to tasks in which the output is not a tree, but for example another sequence of words. Examples of such tasks are language modeling (predicting the next word), spelling correction (converting a distorted sequence to a “clean” sequence), or translation. In the section *Memory-based language modeling and construction grammar* we discuss our memory-based approach to these types of text-to-text processing tasks, and argue that examples in these tasks do not require explicit hierarchical structure, and when some hierarchical structure is needed (e.g. in generating translations), this follows implicitly from the overlap between found examples.

### Memory-based morpho-phonology

Daelemans, Gillis, and Durieux (1994) present a memory-based account of stress assignment to Dutch simplex words. If one would follow the arguments of the then-current principles & parameters-based account of stress assignment applicable to Dutch (Dresher & Kaye, 1990), there would on the one hand be a rule set that determines “regular” stress assignment, while on the other hand there would exist exceptions to the rules, invoked by lexical markings. The rules are based on a notion of syllable weights of the last three syllables, where the weight of each syllable is in turn determined by its rhyme (nucleus and coda); the five-valued weight scale can range from superlight rhymes containing only a schwa, to superheavy rhymes with VCC or VVC structure; an integer value between 1 and 5 represents the weight, as illustrated for a three-syllable Dutch word in Table 1.

Through comparative experiments, Daelemans et al. (1994) show that when words are represented by the weight of the rhymes of their last three syllables, a memory-based learner is able to predict the stress of the “regular” cases well (e.g. the regular penultimate stress pattern is predicted with 93.6% accuracy), while performing relatively weak on other

Table 1: Three representations of the three-syllabic Dutch word agenda, pronounced /a-’gEn-da/, with primary stress on the penultimate (second) syllable, and the percentage of accurately predicted stress assignments for unseen words

Encoding	Syllable			Accuracy (%)
	1	2	3	
Dresher & Kaye (1990) weights	2	3	2	81.2
Rhymes (nucleus and coda)	a –	E n	a –	88.1
Syllables (onset, nucleus and coda)	– a –	g E n	d a –	88.8

cases (final stress 74.9%, and antepenultimate stress 53.2%). When instead the encoding is changed into the actual phonemic content of the nucleus and the coda, performance on cases outside the “regular” set improves (to 87.9% on final stress, and to 61.8% on antepenultimate stress), indicating that the theory’s reliance on an abstracted syllable weight was in fact hiding useful information. In a third experiment, the identity of the onset phonemes was also included, leading to a further increase in performance, and an overall best accuracy of 88.8% correctly assigned stress to unseen test words.

### Memory-based language modeling and construction grammar

Natural language processing models and systems typically employ abstract linguistic representations (syntactic, semantic, or pragmatic) as intermediate working units. Memory-based models enable asking the question whether we can do without them, since any invented intermediate structure is always implicitly encoded somehow in the words at the surface, and the way they are ordered, and memory-based models may be capable of capturing the knowledge that is usually considered to be necessary, in an implicit way, so that they do not need to be explicitly computed.

Classes of natural processing tasks in which this question can be investigated in extremis are processes in which form is mapped to form, i.e., in which neither the input nor the output contains abstract elements to begin with, such as machine translation. Many current machine translation tools, such as the open source Moses toolkit (Koehn et al., 2007), indeed implement a direct mapping of source to target text, leaving all of syntax and semantics implicit; they hide in the form of statistical translation models between collocationally strong phrases, and of statistical language models of the target language. Our take on this problem involves using context on the source side, and using memory-based classification as a translation model (Van Gompel, Van den Bosch, & Berck, 2009).

The model presented in (Van Gompel et al., 2009) uses the phrase alignments computed in Moses (Koehn et al., 2007), producing pairs of word  $n$ -grams on the source and target side

of translation that display a strong mutual conditional probability. As visualized in Figure 2, the fully automatic procedure discovers alignments between pairs such as the English sequence ‘wrongly convicted’ and the French sequence ‘condamné a tort’, which, besides contextually appropriate translations of each other, are both strong collocations, and can both be seen as strong lexicalized constructions in their respective languages.

In the field of statistical machine translation, an interesting development with respect to our discussion on the topic of assuming hierarchical structures in the memorized examples, is that the Moses approach of building phrase tables of pairs of flat word  $n$ -grams has recently been challenged by an approach that assumes phrases to be hierarchical, i.e. phrases can contain sub-phrases (Chiang, 2007). The hierarchical approach has met with some success, although it has not been shown to outperform the Moses approach across the board.

There is an encouraging number of recent studies that attempt to link statistical and memory-based models of language that focus on discovering strong  $n$ -grams (for phrase-based statistical machine translation or for statistical language modeling) to the concept of constructions and to the question to what extent human language users exploit constructions. To mention two, we note that Mos, Van den Bosch, and Berck (To appear) have reported that a memory-based language model shows a reasonable correlation with unit segmentations that test subjects generate in a sentence copy task; the model implicitly captures several strong complex lexical items (constructions), although it fails to capture long-distance dependencies, a common issue with local  $n$ -gram based statistical models. In a related study, (Arnon & Snider, 2010) show that subjects are sensitive to the frequency (a rough approximation of collocational strength) of four-word  $n$ -grams such as ‘don’t have to worry’, which are processed faster when they are more frequent. Their argument is again focused on the question whether strong subsequences need to have linguistic structure that assume hierarchy, or could simply be taken to be flat  $n$ -grams — it is exactly this question that we aim to explore further in our work with memory-based language processing models.

### Acknowledgements

The authors wish to thank Steven Gillis, Gert Durieux, Maria Mos, Maarten van Gompel, and Peter Berck for their contributions to work summarized in this extended abstract, and to Royal Skousen, Harald Baayen, and Emmanuel Keuleers for valuable suggestions and discussions.

### References

- Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67–82.
- Bod, R. (1995). *Enriching linguistics with statistics: Performance models of natural language*. Unpublished doctoral dissertation, ILLC, Universiteit van Amsterdam, Amsterdam, The Netherlands.

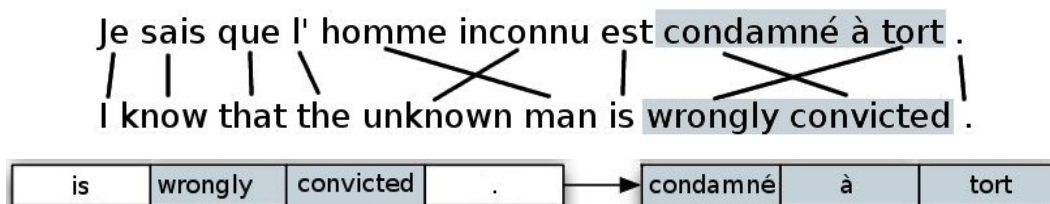


Figure 2: *Above*: A word-aligned sentence pair, with a hypothetical extractable phrase marked. *Below*: A training instance with source-side context for the marked extractable phrase, English to French. Left of the arrow is the feature vector, the class is on the right.

- Bod, R., Scha, R., & Sima'an, K. (Eds.). (2003). *Data-oriented parsing*. CSLI Publications.
- Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2), 201–228.
- Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13, 21–27.
- Croft, W., & Cruse, A. (2003). *Cognitive linguistics*. Cambridge: Cambridge University Press.
- Daelemans, W. (2002). A comparison of analogical modeling to memory-based language processing. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical modeling*. Amsterdam, The Netherlands: John Benjamins.
- Daelemans, W., Gillis, S., & Durieux, G. (1994). The acquisition of stress: a data-oriented approach. *Computational Linguistics*, 20(3), 421–451.
- Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press.
- Daelemans, W., Van den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning, Special issue on Natural Language Learning*, 34, 11–41.
- Dresher, E., & Kaye, J. (1990). A computational learning model for metrical phonology. *Cognition*, 32(2), 137–195.
- Eddington, D. (2002). A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical modeling: An exemplar-based approach to language*. Amsterdam, The Netherlands: John Benjamins.
- Estes, W. K. (1994). *Classification and cognition* (Vol. 22). New York: Oxford University Press.
- Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press, USA.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics.
- Langlais, P., Yvon, F., & Zweigenbaum, P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (eacl-09)* (pp. 487–495). Athens, Greece.
- Lepage, Y., & Shin-ichi, A. (1996). Saussurian analogy: a theoretical account and its application. In *Proceedings of the 16th International Conference on Computational Linguistics, coling-96, copenhagen, denmark* (pp. 717–722).
- Mos, M., Van den Bosch, A., & Berck, P. (To appear). The predictive value of word-level perplexity in human sentence processing: A case study on fixed adjective-preposition constructions in dutch. In D. Divjak & S. Gries (Eds.), *Corpus and cognition: Converging and diverging evidence*. Berlin: Mouton De Gruyter.
- Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 15, 39–57.
- Post, M., & Gildea, D. (2009). Bayesian learning of a tree substitution grammar. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 45–48).
- Skousen, R. (1989). *Analogical modeling of language*. Dordrecht: Kluwer Academic Publishers.
- Skousen, R., Lonsdale, D., & Parkinson, D. B. (Eds.). (2002). *Analogical modeling: An exemplar-based approach to language*. Amsterdam, The Netherlands: John Benjamins.
- Smith, E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
- Smith, L., & Samuelson, L. (1997). Perceiving and remembering: Category stability, variability, and development. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 161–195). Cambridge: Cambridge University Press.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
- Van Gompel, M., Van den Bosch, A., & Berck, P. (2009). Extending memory-based machine translation to phrases. In M. Forcada & A. Way (Eds.), *Proceedings of the Third Workshop on Example-Based Machine Translation* (pp. 79–86). Dublin, Ireland.
- Yvon, F., & Stroppa, N. (2007). Proportions in the lexicon: (re) Discovering paradigms. *Lingue e linguaggio*, 2, 201–226.