Article

Language *and* Speech

Implicit Schemata and Categories in Memory-based Language Processing

Language *and* Speech
56(3) 309–328

DOI: 10.1177/0023830913484902
las.sagepub.com
SAGE

Antal van den Bosch
Centre for Language Studies, Radboud University Nijmegen, The Netherlands

Walter Daelemans
Computational Linguistics and Psycholinguistics Research Centre, University of Antwerp, 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 as a model for human language processing because it does not make any assumptions about the way abstractions are shaped, nor any a priori distinction between regular and exceptional exemplars, allowing it to explain fluidity of linguistic categories, and both regularization and irregularization in processing. Schema-like behaviour and the emergence of categories can be explained in MBLP as by-products of analogical reasoning over exemplars in memory. We focus on the reliance of MBLP on local (versus global) estimation, which is a relatively poorly understood but unique characteristic that separates the memory-based approach from globally abstracting approaches in how the model deals with redundancy and parsimony. We compare our model to related analogy-based methods, as well as to example-based frameworks that assume some systemic form of abstraction.

Keywords
Language acquisition and processing, memory-based learning

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. Our specific operationalization of these ideas is relatively recent (Daelemans & Van den Bosch, 2005), but the ideas have been around for a long time.

Corresponding author:
Antal van den Bosch, Centre for Language Studies, CIW/BC, Faculty of Arts, Radboud University Nijmegen, P.O. Box 90153, NL-5000 LE Nijmegen, The Netherlands.
Email: a.vandenbosch@let.ru.nl

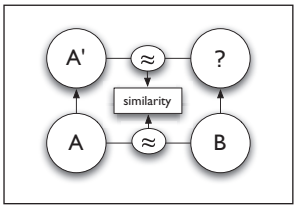

Figure 1. Saussurean analogy. Given the proportional analogy A:A'::B:B', with B' missing, we need to find a B' that fits the analogy.

We see MBLP as the implementation of the analogical, example-based strand of linguistic theories developed throughout the 20th century, from the neo-grammarian notion of *Analogiebildung* (Paul, 1880) and Saussurean analogical reasoning (De Saussure, 1916) to example-based models of human language processing. Figure 1 illustrates the Saussurean analogical proportion that holds between sequences A and B (e.g. orthographic words), and their mappings A' and B' at another linguistic level (e.g. phonemic words). If B' is not known in advance, it is the *quatrième proportionelle* that can be computed on the basis of the other three parts of the four-way proportion, which states that A:B::A':B', meaning that that A' and B' stand in the same similarity relation as A and B, and that B' is therefore as similar to A', as B is similar to A.

In most cases in natural language processing it is impossible to find an analogical proportion that predicts B' fully, due to the sparseness of long subsequences in language. Hence, except for a particular branch of work on full analogical proportions on which we expand in Section 5.1, a full Saussurean analogical reasoning process is typically decomposed 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'. This decomposition and 'subsequencing' approach is the basic template for the majority of present-day mainstream statistical natural language processing algorithms, not only for operationalizations of Saussurean analogy. Memory-based learning is one of the few models within the multitude of available approaches that bases its decisions on extrapolations from examples, rather than on amalgamate models that abstract away from single examples.

We can take the Saussurean analogical proportion as the framework for an analogical reasoning process fit for computer implementation. As Figure 1 visualizes, generating B' is essentially the task of finding a B' that best fits the Saussurean analogical proportion among many possible candidates: we need to find a B' that is as similar to A' as A is to B. Broadly speaking, this is the task that we nowadays see tackled by probabilistic models and supervised machine-learning algorithms, but a key ingredient is lost: that the analogical proportions are drawn between individual instances of sequences. Instead, most machine-learning approaches and all probabilistic models compare B to a numeric model that represents global statistics gathered over many individual examples; no analogy is drawn from single examples.

In contrast, MBLP is a member of a class of machine-learning algorithms that runs the Saussurean analogy-based reasoning process whenever a new B is presented. This class of algorithms finds its computational basis in the classic *k*-nearest neighbour 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'. With *k* set to higher values, the *k*-nearest neighbours to B are retrieved, and some voting procedure (such as majority voting) determines which value is copied to B'.

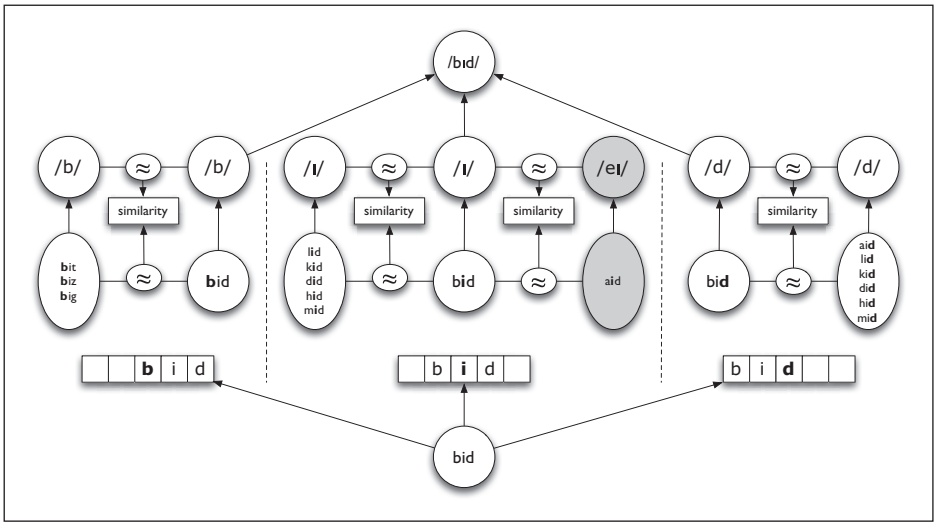

Figure 2. Decomposed analogical reasoning over the letter-phoneme correspondence of each of the three letters of 'bid'. The ovals contain nearest neighbours found. All neighbours are 'friendly' except the word 'aid', for the phonemization of the letter 'i'; this unfriendly neighbour is marked in grey.

A common variant of the pure Saussurean proportional analogical reasoning that operates on sequences of any length is the decomposition of variable-length sequences into fixed-length subsequences. Limiting and fixing the representation space often turns out to be necessary to allow for sufficiently reliable analogical reasoning: the limitation leads to better and less sparse matches.

Figure 2 illustrates this decomposition: the phonemization of the word 'bid' is decomposed in three letter-phoneme mappings, each focusing on one of the three letters. The first of these mappings matches the letter 'b' with its empty left context and 'id' as its right context to subsequences in memory, finding the three most similar neighbours derived from the words 'bit', 'biz' and 'big'. All three subsequences point to /b/ as the contextually appropriate phonemic mapping of the 'b'. A minor disagreement arises in the mapping of the middle letter 'i'. One of its most similar neighbours, derived from the word 'aid', maps to /eɪ/. This is offset against five neighbours mapping to (voting for) the phonemization /ɪ/; a simple way to solve this conflict is to let the majority win. Finally, the individual phonemes produced by the three analogical reasoning steps are concatenated to form the resulting phoneme sequence /bɪd/.

A key element in performing analogical reasoning is the similarity function used to establish a graded notion of distance (the reverse of similarity) between any pair of examples, ranging from zero with two identical examples, to a maximum value for two examples that have no feature values in common. The distance function is a sum of pairwise value distances per feature, possibly weighted by the relative importance of the feature estimated through information-theoretic metrics, so that the nearest neighbour to a new example will be an example with the least mismatches on the most important features. In the case of positional features, such as the windows in Figure 2, it is more important for a nearest neighbour to match on the middle letter than on any of the context letters; 'bed' is not a good example of the pronunciation of the 'i' in 'bid', while 'hip' would be: this means that the similarity function should take into account that a match on the middle letter should weigh more heavily than a match on both the left and right context letter, which indeed is what the

information-theoretic metrics will estimate. The nearest neighbourhood may also comprise a set of equally similar examples, such as 'lid', 'kid', 'did', 'hid', 'mid' and 'aid', which all mismatch in the letter left to the centre 'i' in 'bid'). When lacking an ideal exact match, the search for a nearest neighbour will back off to less ideal matches. For technical details on the implementation of the similarity function in MBLP, the reader is referred to Daelemans and Van den Bosch (2005).

In this article we provide two perspectives that highlight unique properties of the memory-based approach. Firstly, in Section 2, we take a spatial perspective and characterize the type of class spaces one finds in natural language processing. These spaces turn out to be highly disjunct: when looking for nearest neighbours around a single example B, the nearest example A often has an A' that is not at all as similar to B' as A is to B. This fact stands in conflict with the Saussurean analogy principle, but we explain why the situation is not that bad.

In Section 3 we explain how a memory-based classifier operates locally, in contrast to essentially the rest of the machine-learning and probabilistic natural language processing algorithms. This local classification is in fact a process called *selective abduction*, which is how the Saussurean analogy principle can also be explained. We argue that the flexibility offered by local classification provides valuable computational advantages that no global model can offer, such as incrementality and decrementality, which can be used for modelling online processing, learning and forgetting.

As a computational model of human language processing, MBLP can be seen from the perspectives of several strands of research in psychology on categorization, episodic memory and global memory matching. MBLP can also be connected to ideas in usage-based linguistics. We discuss these relations in Section 4.

The memory-based learning approach has some computational nearest neighbours that we discuss in Section 5: Skousen's Analogical Modeling, the work of a group of researchers we refer to as the French Analogical Proportionists, and data-oriented parsing (DOP). In Section 6 we zoom in on two widely different language processing tasks, stress assignment in Dutch simplex words and translation, to highlight particular aspects of memory-based learning. We conclude in Section 7 by summarizing our arguments.

2 How friendly are linguistic neighbourhoods?

When given a corpus of annotated examples of a natural language processing task, the corpus can be transformed into a set of mappings of the A-A' type. As already pointed out in the introduction, it is common to constrain the format of both A and A', for example to fixed-length subsequences of symbols. The most common template, known as *windowing*, takes a complete mapping of two structures, for example a sentence and a non-nested bracketing of that sentence into syntactic chunks, or a word and its phonemic transcription, or a sentence in a source language and its translation in another language, and transforms this into a number of mappings of fixed-length subsequences of words, each focusing around one word or letter. An example of this process is visualized in Figure 3. On the left, the figure shows an English sentence and its Dutch translation along with statistically established word alignments. On the right-hand side of the figure, each of the four word alignments is now the focus of a mapping between a fixed-width trigram of source tokens to a fixed-width trigram of target tokens (Van den Bosch & Berck, 2009).

Windowing ensures that analogical reasoning is always applied to subsequences of the same length. Comparing same-length subsequences offers significant computational advantages over comparing subsequences of different lengths.

While fixed-width windows fit sequential tasks such as grapheme-phoneme conversion and part-of-speech tagging quite well, it offers at best only a half-way solution in structured learning

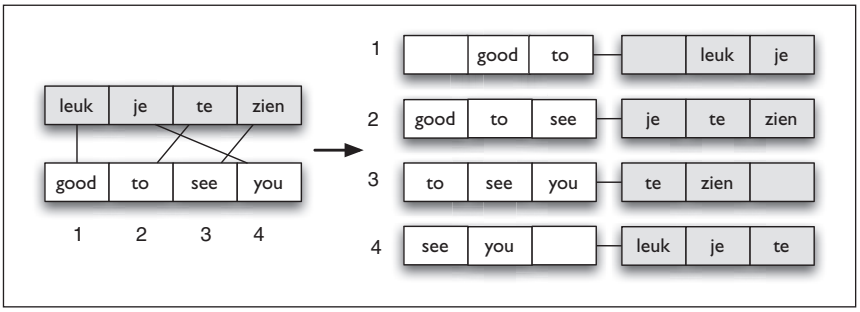

Figure 3. Windowing: the transformation of the mapping of the English input sequence 'good to see you' to the Dutch output sequence 'leuk je te zien', to four mappings of trigrams of input tokens to trigrams of output tokens, where the middle tokens align.

tasks, in which the output is not necessarily a sequence with the same length as the input sequence. The output can have a different length and order such as a translation, or may be an entirely different type of structure altogether, such as a parse tree. In both cases, a strong inference process involving search in considerably large spaces, and possibly domain knowledge, is needed to transform the outcome of local analogical reasoning steps, such as the one in the right-hand side of Figure 3, to the desired outcome. In the case of Figure 3, the output representing the correct Dutch word order may be reconstructed by searching for the best overlap between predicted trigrams. Two starting points that discuss this issue and present classes of solutions are Lafferty, McCallum, and Pereira (2001) introducing Conditional Random Fields, and Punyakanok, Roth, Yih, and Zimak (2005) discussing learning and inference using constraints on the output solution space. Although an important part of natural language processing technology, we ignore these structured output issues for now and concentrate on the first step again: local analogical reasoning.

When storing large amounts of windowed A-A' examples of a particular natural language processing task in a computer memory, it is possible to use this memory to handle new processing requests involving B as input, and producing B' as output. It is also possible to compare all stored examples with each other to learn some general facts about the space they occupy. This space has as many dimensions as the examples have features. The windowed examples of Figure 3 have three positional features, in which several values have been seen occurring in the examples presented as training data. In other words, the four examples in the figure occupy four points in a three-dimensional space, where each dimension is symbolic, representing an unordered set of symbols. Each example in this space will have one or more nearest neighbours at some distance (e.g. at a distance of one differing value). Assumedly, following De Saussure, close nearest neighbours will tend to map to the same output (we refer to such cases as 'friendly neighbours'), while at the same time there must be points in the space of which the nearest neighbour maps to a different output, as rival subspaces of different outcomes must border each other somewhere.

In earlier work (Daelemans & Van den Bosch, 2005) we explored this issue and traced the nearest neighbours for all examples A in memory, for four natural language processing tasks represented as windowed or otherwise fixed-length examples, ranging from morpho-phonological tasks to syntactic and information extraction tasks. We found that between about 10% and 20% of all examples of the tasks did not have a single friendly neighbour. This number in itself does not say yet whether this 10–20% of examples is positioned at the border of a large subspace, or whether

Table 1. Numbers of different possible outcomes, families of nearest neighbours and average family size of four natural language processing tasks.

Task	Number of outcomes	Number of families	Average number of family members	% Examples w/ unfriendly neighbours	% Generalization error
German noun pluralization	8	17,49	7.2	15.3	6.0
Dutch diminutive inflection	5	233	12.9	9.1	2.4
English PP attachment	2	3,613	5.8	16.5	19.3
English base phrase chunking	22	17,984	11.8	14.9	8.1

they form small single-example subspaces surrounded by unfriendly neighbours; in other words, we do not actually know how dispersed the example space is in terms of clusterings of friendly neighbours – are neighbours with single outcomes clustered in a single large subspace, or are they scattered and mixed with other scattered clusters of friendly neighbours with other outcomes? To explore this further, we devised an algorithm (Fambl, see Van den Bosch, 1999, and below) that groups together friendly neighbours iteratively in hyperballs, thus forming 'families' of friendly neighbours that can be likened to explicitly generalized schemata. The second and third columns of Table 1 list the numbers of families and average family size that we encountered this way in four natural language processing tasks (for details, see Daelemans & Van den Bosch, 2005). It is obvious from these numbers that the example spaces of these tasks are highly disjunct with respect to the clusteredness of examples mapping to the same outcome. The average family size, displayed in the fourth column of Table 1, ranges between only about 6 and 13. The members of these families are each other's nearest neighbour, but there are many cases where members of two bordering families are each other's nearest neighbour. The fifth column of Table 1 shows the percentage of examples that have a differently labelled ('unfriendly') neighbour as their immediate nearest neighbour; for the four example tasks, this percentage ranges from 9% to 17%.

One implication of this finding is that the Saussurean analogy principle must fail at the many boundaries in space where nearest neighbours map to different outcomes. A sensible processing system should be somehow aware of these boundaries, and should not make the error of copying the incorrect A' to a B that stands on the wrong side of the border. Memory-based learning does not explicitly draw global decision boundaries: it only assumes an implicit division of the space. When the analogical reasoner operates on only the single nearest neighbour, with $k = 1$, the space is implicitly divided into a so-called Voronoi tessellation of the kind displayed in Figure 4, exemplifying a two-class space where the classes are black and white, and where the six displayed examples are characterized by their coordinates in the space. Each tile is occupied by one example.

Since the memory-based classifier is not aware of boundaries that separate areas with examples mapping to different output symbols, its analogical reasoning is likely to produce errors. When 9–17% of all memory examples in the four linguistic spaces listed in Table 1 have an 'unfriendly' nearest neighbour, we may assume that the same percentage of new examples will have 'unfriendly' nearest neighbours as well. If this would be the case (and if $k = 1$), analogical reasoning can be expected to make about the same number of errors. In practice, however, it often makes fewer errors on unseen data than that, as a comparison of the fifth and sixth columns of Table 1 shows; the sixth column displays the average generalization error (a percentage between 0 and 100) as reported by Daelemans and Van den Bosch (2005). For one, this is because any unseen data, drawn from the same population of examples that the training set was drawn from, will contain examples that are very similar to, or even duplicates of examples in memory, so that they find a nearest

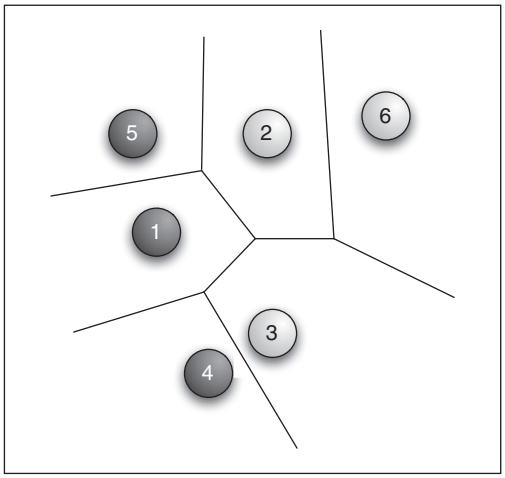

Figure 4. Voronoi tessellation of a two-class two-dimensional space with $k = 1$.

neighbour well inside friendly neighbourhoods. Examples are not necessarily identical; it is sufficient that their representations are the same to act as a duplicate. We know empirically that exact or very close matches tend to yield accurate analogical reasoning (Daelemans & Van den Bosch, 2005) if the representation is chosen well.

Yet, we do see analogical reasoning fail in some cases, reasoning across boundaries it should not cross. Performance figures, such as generalization accuracy on test data, of memory-based learning can in fact be seen to quantify the degree to which the Saussurean analogy principle explains unseen data. Regardless of the task, we have never observed a 100% fit. We did, however, observe the following after closer analysis of the predictions of our memory-based classifier.[1]

1. Due to the total recall of all training examples that the simplest version of our software exhibits, any duplicate example that recurs in new data will always be remembered, and the analogical reasoning generally generates accurate predictions in these cases of mere recall.[2] This 'lookup' property, simple as it is, offers a strong backbone for MBLP. The implied golden rule is to always remember all training examples, as they may recur in the future, as do all patterns in language in due course.
2. This principle even holds for examples that seem to be atypical or rare, or simply infrequent. In case of word features, for instance, many examples will contain values residing in the long Zipfian tail (Zipf, 1935). We have shown in earlier work (Daelemans & Van den Bosch, 2005; Daelemans, Van den Bosch, & Zavrel, 1999) that there is rarely any positive effect gained from ignoring training examples estimated to be infrequent, atypical or bad, and that we in fact observe negative effects when leaving out examples, regardless of the criterion by which they are ignored. Rare examples can be friendly neighbours too.
3. However, we have observed in several experiments that random removal (i.e. forgetting) of examples from memory can be done without harming generalization performance significantly. Examples include up to 80% forgetting with English prepositional attachment, and 90% forgetting with Dutch diminutive inflection generation (Daelemans & Van den Bosch, 2005, p. 129). It appears that most of our training sets contain more than a sufficient

number of examples (such as duplicate example tokens, or very similar examples labelled with the same outcome) that can be removed without changing the behaviour of the classifier. The Fambl approach described earlier (Van den Bosch, 1999) is a mechanism that implements this shallow bottom-up generalization of homogeneous instances into generalized examples that can be likened to explicit schemata.

In sum, through experimentation we have shown that although Saussurean analogical reasoning, a form of selective abductive reasoning, has no built-in safety measures to avoid erroneous outcomes, MBLP can offer good performance on unseen data in many natural language processing tasks by using the principle directly, offering the best possible performance when all of the training data is retained in memory. In other work we and others have shown that memory-based learning offers competitive state-of-the-art performance, so that it can be used in most practical natural language processing situations in lieu of any other state-of-the-art machine-learning algorithm (for an overview, see Daelemans and Van den Bosch, 2009).

3 Local versus global modelling

The unique characteristic of memory-based classification, or *k*-nearest neighbour classification, is that it does not make use of the same single model at every classification, like virtually every other machine-learning algorithm does. Instead, it builds temporary and different models each time it receives a classification request. This characteristic has lent *k*-nearest neighbour classification the moniker *lazy learning*. No effort is invested in abstracting classification knowledge from the available training data (or, to put it differently, no distinct learning phase takes place); instead, the training data is kept in memory as is; only when needed a small local model is generated from the data that is forgotten again immediately after usage.

Such a temporary model takes the form of a hyperball that extends into the feature space introduced in the previous section, up to the radius at which the designated number of *k*-nearest neighbours or nearest distances (i.e. distances at which equally distant nearest neighbours are found) is within the ball. This is illustrated in Figure 5. As an alternative to *k*, a fixed radius *d* can be set that fixes the size of the hyperball; this is usually referred to as a Parzen window (Parzen, 1962). With *k*, the radius of the local hyperball varies; with a Parzen window, the actual number of nearest neighbours captured in the fixed-sized ball varies. There are no general rules for setting *k* or *d* or whether either of the two is to be preferred.

Subsequently, the contents of the hyperball are the basic ingredients of the temporary model. The output symbols (outcomes) contained in it can each be seen as a vote for those symbols; when $k > 1$, some voting system is needed on top of the basic Saussurean analogy to boil the votes of all the nearest neighbours down to a single outcome. Several mechanisms have been proposed for voting. For example, it is typical to assign lower weights to votes of more distant nearest neighbours in the hyperball, transforming their distance to a weight according to some function, such as inversing it (cf. Daelemans & Van den Bosch, 2005). The output symbol that receives the largest aggregate of weighted votes, summed over the votes of all nearest neighbours, is produced as the outcome. Alternatively, the full distribution of vote weights can be generated as the outcome. When the weights are normalized to add up to 1.0, this distribution can be taken as a local probability distribution, making it suitable for further processing in a probabilistic framework.

After the analogical reasoning has been performed, the local hyperball and its contents are forgotten, and the system switches to a waiting state ready to handle a new reasoning request. This

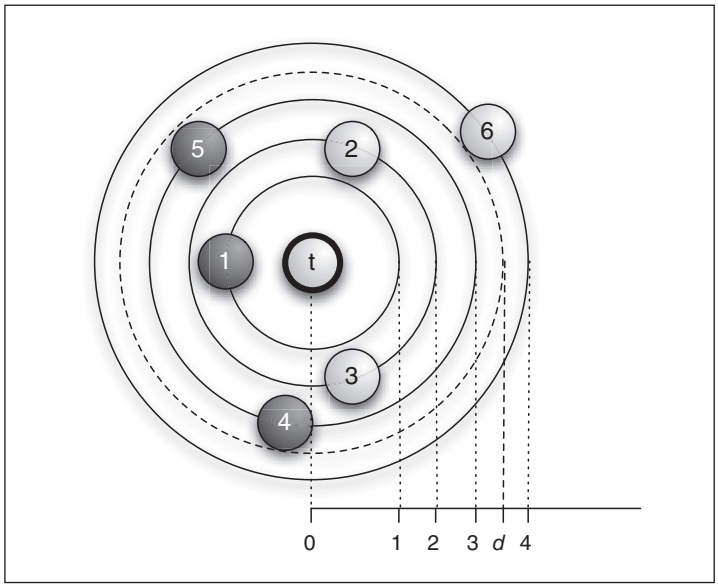

Figure 5. Radii of four nearest distances at which nearest neighbours are found of the new example *t*; also, a Parzen window at distance *d* is drawn.

type of processing can be likened to an experienced medical doctor, waiting for patients and diagnosing them as they come in, on the basis of their symptoms. This is actually the application for which the nearest neighbour algorithm was first proposed (Fix & Hodges, 1951). The doctor has seen patients with certain diseases exhibiting a certain set of symptoms. Observing a sufficiently similar set of symptoms in a new patient triggers the doctor to hypothesize, or *abduce*, that the patient may be suffering from the same disease as earlier patients who shared most of the symptoms. In this process, the doctor draws on his memory of patients, but his diagnosis is not based on a model of all the patients he ever saw; rather, he performs a *selective abduction* on the basis of the most similar patients. Thus, memory-based reasoning follows a process of selective abduction: not induction, which assumes a separate learning phase in which rules are induced from examples, or deduction, which assumes knowledge of all necessary causative rules (if a patient has this disease, he is likely to have this set of symptoms). Like induction, selective abduction can produce faulty outcomes, as we have already seen in the previous section. Yet, it is a flexible, lightweight method of reasoning because it does not rely on an expensive error-prone learning process (learning rules from examples by induction) or on a costly and brittle resource (flawless high-coverage causative rules for deduction).

The fact that MBLP does not rely on a single learned model for all of its predictions has a clear advantage: it naturally allows incremental learning and decremental learning, or forgetting. The addition and removal of examples from memory can be done at any point in time, and does not imply a costly and cognitively implausible recomputation of a global model.

Beyond the mere removal and addition of individual examples, a related property of MBLP is that individual examples can be given a unique weight that can influence the strength of the vote that this example casts when involved as a nearest neighbour in an analogical reasoning process. In this way activation, inhibition and gradual forgetting can be modelled, which offers possibilities for computational psycholinguistic modelling.

4 Memory-based language processing as a model of human memory and language processing

MBLP can be understood as an implementation of exemplar-based models of human memory, and can be connected to psychological work in this area – which tends to use language processing tasks such as word recognition as favourite benchmarks. In particular we see MBLP as an implementation of episodic memory (Goldinger, 1998; Hintzmann, 1988; Logan, 1988; Raaijmakers & Shiffrin, 1980).

In their review of episodic memory models that rely on the matching of new input to exemplars stored in memory, which they refer to as *global matching models*, Clark and Gronlund (1996) define these models by two key characteristics: '(1) recognition is based solely on familiarity due to a match of test items to memory at a global level, and (2) multiple cues are combined interactively' (Clark & Gronlund, 1996, p. 37). We see these characteristics reflected in MBLP if we translate the terminology. When Clark and Gronlund refer to global matching, they mean the comparison of a test item to all exemplars in memory (rather than the retrieval of one particular exemplar). The interactive combination of multiple cues can be translated to the computation of a function over a vector of features. Their models, like MBLP, can account for both the retrieval of exact ('intact') matches from memory, and the matching with similar but non-identical exemplars.

The work of Lewis, Vasishth, and Van Dyke (2006) on cognitively plausible models of memory in sentence comprehension is relevant with respect to MBLP as it points at the implausibly slow memory retrieval of serial order, which takes hundreds of milliseconds (McElree, 2006), versus the fast comparison of cues to items in memory (80–90 ms) that is sufficiently fast to explain sentence comprehension (250–300 ms per word). MBLP models the latter type of processing: as a sentence processing model it can be seen as a processor of a sequence of cues, where each cue represents a local subsequence of the sentence. While Lewis et al. (2006) equate cues with words, we believe it is reasonable to assume that the cues and exemplars in human sentence processing are variable-width local sequences ranging from letter *n*-grams to a handful of words, representing a local context that offers sufficient information to find similar nearest neighbours in memory, and use analogical reasoning over them to trigger a correct response.

Exemplar-based models also play a pivotal role in psychological studies of human categorization, and have been argued to produce a generally good fit of human behaviour and errors (Estes, 1994; Nosofsky, 1986; Smith & Medin, 1981). 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. Even the very assumption of fixed, permanent categories (however represented) has come under fire by theories favouring 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 (Croft & Cruse, 2003; Smith & Samuelson, 1997). This type of context-dependent, memory-based category emergence fits MBLP well.

One recent approach to linguistics, usage-based models of language, proposed 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, and the scheme-like behaviour emerges from the exemplar-based processing. In this sense, MBLP is less redundant than cognitive linguistics. Other aspects of cognitive linguistics, such as the importance of frequency, and experience-based language acquisition (Tomasello, 2003), fit MBLP naturally as well. As far as frequency is concerned, many experiments with MBLP have focused on type frequencies, rather than token frequencies, the latter playing an important role in cognitive linguistics, for instance to model entrenchment of exemplars and schemata. MBLP can accommodate token frequencies by representing them (or a log normalization) as exemplar weights. In the TiMBL implementation of MBLP, the distance between a new exemplar and a memory exemplar is divided by the weight of the memory exemplar (e.g. representing its token frequency), so that more frequent exemplars are drawn closer in distance to new exemplars (Daelemans & Van den Bosch, 2005). In our experiments on language processing tasks, we have found type frequencies to lead to better generalizations in morphological tasks, and token frequencies to sometimes play a role in syntactic tasks.

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 modelling) to the concept of constructions and to the question of to what extent human language users exploit constructions. Wiechmann (2009) focuses on English relative clause constructions, and proposes a framework that combines construction grammar theory with example-based processing. He shows that his model exhibits high degrees of compatibility both with quantitative corpus data and experimental data obtained with humans. Recently, Mos, Van den Bosch, and Berck (2012) reported that a memory-based language model shows a reasonable correlation, explaining over 25% of the variance in segmentations that test subjects generate in a sentence copy task. The model implicitly captures several strong units, but fails to capture long-distance dependencies, a common issue with local *n*-gram-based statistical models.

Bannard and Matthews (2008) show with sentence-repetition tests that children repeated frequent sequences significantly more correctly than infrequent sequences. Inspired by this study, Arnon and Snider (2010) show that subjects are sensitive to the frequency of four-word *n*-grams such as ‘don’t have to worry’, which are processed faster when they are more frequent. Arnon and Cohen (this issue) report on experiments of which the results indicate that phonetic duration is reduced in multi-word sequences with a higher frequency, regardless of the syntactic boundaries these sequences cross. Furthermore, the phonetic duration effects cannot be explained by the frequencies of the individual words or subsequences. The discussion in both studies homes in on the question whether strong sequences 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 MBLP models.

In this work we will need to address the challenge of Baayen, Hendrix, and Ramscar (this volume), who propose a very compact association model between letter *n*-grams and meaning nodes based on naïve discriminative learning that is able to explain the same frequency effects in comprehension tests that *n*-gram frequencies from a corpus can explain (Arnon & Snider, 2010). Yet, representing hundreds of millions of *n*-gram counts represents substantially larger storage costs, which Baayen et al. (this volume) consider cognitively less plausible than their more parsimonious alternative. It remains to be seen how the naïve discriminative learning framework would scale up to representing not under 8000 ‘meanings’ (morphemes) selected to strictly fit Arnon and Snider’s

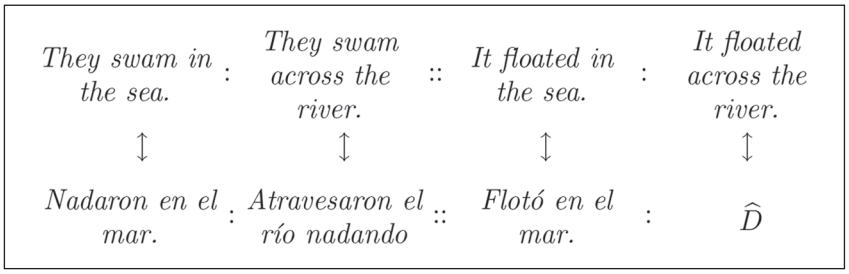

Figure 6. A parallel proportional analogy between four sentences in English, and four sentences in Spanish of which one is missing; the system of Lepage and Denoual infers the missing translation by analogy from the known sentences. Taken from Lepage and Denoual (2005).

(2010) data, but the hundreds of thousands of unique morphemes present in the 20 million word corpus from which Arnon and Snider derived their *n*-gram frequencies.

5 Related computational approaches

5.1 Analogical modelling and analogical proportions

There is a close relation between MBLP as described so far in this article, and two schools of analogical modelling 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 (e.g. Langlais, Yvon, & Zweigenbaum, 2009; Lepage & Shin-ichi, 1995; Yvon & Stroppa, 2007). The relation of MBLP with Skousen's analogical modelling work has been discussed earlier by Daelemans (2002). The analogical modelling approach is memory-based in that all available individual examples are used in extrapolating to the solution for a new input. The set of nearest neighbours considered for extrapolation is not constrained to a k or a hyperball radius, but by a match on any subset of features with the new example to be processed. As the algorithm searches for all examples that share all combinatorial subsets of input features, it is exponential in the number of features, which makes the approach impractical for problems with many features. The approach has been applied to different problems in language processing, mainly in the phonology and morphology domains. Empirical comparisons have never shown important accuracy or output differences between the two approaches (Eddington 2002; Krott, Schreuder, & Baayen, 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 – not our simplified version of windowed input and output subsequences (as in Figure 3). A prototypical example is the work on analogical machine translation by Lepage and Denoual (2005) – the article's title claims that the approach is the 'purest ever example-based machine translation' – which makes use of the parallels between two full analogies between four sentences in two languages, as visualized in Figure 6. No assumptions are made except that sentences are sequences of characters, and can be sliced at any character position to construct a new translation from analogy. Although it is obvious that this system needs considerable numbers of training examples to work, it is shown to perform quite well in the 'basic travelling expressions' domain, in which many formulaic patterns recur.

One difference between the analogical proportionist approach and the memory-based approach concerns the degree to which analogical proportions are computed: on full sequences or on subsequences. Operating on full sequences leans on the intuition that correct outputs can sometimes be generated by combining not more than two partial solutions (at the cost of needing considerable amounts of training data, and at the risk of being confronted with data sparseness), while operating on subsequences leans more on the intuition that with a considerably lower data sparseness problem, very accurate local analogies can be drawn, that later can simply be concatenated to produce a full outcome (e.g. a full translation of a sentence, or a full word pronunciation). If the latter is not 'simply' the case, it is possible to complement good local solutions with a powerful constraint-based search or inference method (Canisius, Van den Bosch, & Daelemans, 2006; Punyakanok et al., 2005).

5.2 Example-driven stochastic models

A looser family relationship exists between analogical methods and example-driven stochastic models such as DOP (Beekhuizen, Bod, & Zuidema, this volume; Scha, Bod, & Sima'an, 2003), and the Bayesian approach described by O'Donnell, Snedeker, Tenenbaum, and Goodman (2011). Data-oriented models of the DOP type are essentially monolithic probabilistic models, as they divide a global probability mass over a large population of labelled tree fragments, which in the original DOP approach is the set of all possible fragments in the treebank (Scha et al., 2003), but which may also be a more parsimonious set of fragments found by a search that starts with long fragments and searches for a minimal description length of the fragment grammar (Beekhuizen et al., this volume). Yet, Scha, Bod, Sima'an and colleagues do stress in their work on DOP the reliance of the method on individual examples. A DOP 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.

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 MBLP model assumes no predefined structure. The DOP approach, the fragment grammar approach of O'Donnell et al. (2011) and also the recent approach proposed by Post and Gildea (2009) imply 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 syntactic tasks. Despite its wide applicability from morpho-phonology to syntacto-semantic 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 modelling (predicting the next word), spelling correction (converting a distorted sequence to a 'clean' sequence) or translation. In Section 6.2 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; when some hierarchical structure is needed (e.g. in generating translations), this follows implicitly from the overlap between found examples.

The crucial difference is that the DOP, O'Donnell and Post and Gildea approaches do not use analogical reasoning, and can therefore not escape from the specific inventory of fragments present in the training data. Creativity at the level of novel combinations is possible in this type of approach, but not productivity through analogy over memory items. The latter type of productivity seems desirable in a psychologically plausible computational model to explain creativity in both regularization and irregularization.

Table 2. Three representations of the three-syllabic Dutch word agenda, pronounced /a-'gɛn-da/, with primary stress on the penultimate syllable, and the percentage of accurately predicted stress assignments to unseen words.

Encoding	Syllable 1	Syllable 2	Syllable 3	Accuracy (%)
Dresher & Kaye (1990) weights	2	3	2	81.2
Rhymes (nucleus and coda)	a-	ɛn	a-	88.1
Syllables (onset, nucleus and coda)	-a-	gɛn	da-	88.8

5.3 Back-off smoothing and decision trees

We have stressed differences of the memory-based approach with other data-driven methods for natural language processing, where the defining characteristic of the former is the reliance on individual examples, not a single model. When we stated there is virtually no other machine-learning algorithm that does this, we skipped over two grey areas where the memory-based approach is similar or even equivalent to abstracting approaches. The first equivalence has been noted by Zavrel and Daelemans (1997), who observe that and explain why a memory-based approach can generate the same predictions as a maximum-entropy classifier with Katz back-off smoothing (Katz, 1987) when trained on the same examples. Under certain conditions, relating to which back-off order is chosen, Katz back-off smoothing is equivalent to a hyperball method that extends into feature space until it encompasses sufficiently matching nearest neighbours.

A second more gradient equivalence can be found between the memory-based approach, and rule-based (e.g. Cohen, 1995) and decision-tree-based (e.g. Quinlan, 1993) methods. The latter methods spend a learning phase on segmenting the example space in sufficiently homogeneous regions in terms of output symbols. With the right parameter settings, rule learners and decision-tree learners can be instructed to segment areas that only contain single examples when needed, and this approaches the situation of a memory-based classifier operating on Voronoi tiles. Daelemans et al. (1999) explored this equivalence, finding indeed that decision-tree learners could be made to classify more like memory-based classifiers if their algorithmic parameters are tuned to fit individual examples more.

6 Case studies in linguistic hypothesis testing through memory-based language processing

6.1 Theory testing in memory-based stress assignment

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 and 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 vowel–consonant–consonant (VCC) or vowel–vowel–consonant (VVC) structure; an integer value between 1 and 5 represents the weight, as illustrated for a three-syllable Dutch word in Table 2.

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

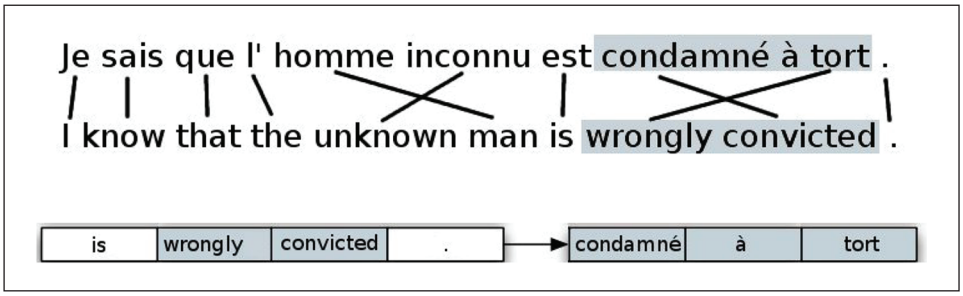

Figure 7. (a) A word-aligned sentence pair, with a hypothetical phrase marked. (b) A training instance with source-side context for the marked phrase, English to French. The input feature vector (left) is associated to the trigram output class (right).

predict the stress of the 'regular' cases well (e.g. the regular penultimate stress pattern is predicted with 93.6% accuracy), while performing relatively weakly on other 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.

6.2 Memory-based translation

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. 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 extremo are processes in which form is mapped to form, in which neither the input nor the output contains abstract elements to begin with, such as 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 statistical translation models between collocationally strong phrases, and of statistical language models of the target language. This reliance of phrasal fragments (here, pairs of aligned fragments) is reminiscent of the fragment-based DOP approach; similar recombination and search procedures are used in the two approaches to produce an output estimated to be maximally likely. Our take on this problem, in contrast, involves analogical reasoning over fragments in their local context (Van Gompel, Van den Bosch, & Berck, 2009), rather than manipulating context-less phrasal fragments.

The model presented by 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 7, 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 Haque, Naskar, Van den Bosch, and Way (2011) a direct comparison is made between the phrasal fragment-based Moses system, and a variant of Moses where the simple phrase lookup procedure is replaced by the translation model of Van Gompel et al. (2009). Instead of looking up all possible candidate phrasal fragment translations in a phrase table given an input phase, analogical reasoning refines this selection by suggesting phrasal fragment translations of local nearest neighbours of the current source-side fragment and its context. They observe that performing analogical reasoning to determine the locally likely phrasal translations of source-language phrases produces significant improvements over the Moses approach in several machine translation benchmark experiments, and never produces a loss in performance (Haque et al., 2011).

7 Discussion

In this contribution we presented MBLP as an approximate implementation of the Saussurean analogy. The full Saussurean analogy, over full sequences, suffers from sparsity. A working model of the Saussurean analogy needs to find analogies in spaces of which the population is still dense enough to find good, 'friendly' neighbours similar to (but not necessarily the same as) the new input. We solve this problem in MBLP by seeking task representations that zoom in on local classifications of (e.g. windowed) subsequences.

The case for MBLP, apart from the fact that it is computationally feasible, extends to a case for a cognitive theory of language processing that is optimally parsimonious in that it assumes that exemplars and analogical reasoning (or episodic memory and global memory matching) are all that is needed for modelling both learning and processing. We translated this to computational models in which the data, and nothing but the data, is the model. In this model, processing happens in the temporary analogical reasoning over examples; in other words, processing lies in the relations temporarily established between exemplars. Linguistic mappings or categories are dynamically construed, and are not permanent, and neither are the rules or schemata that could be used to explain why the memory-based model makes a certain prediction. For that reason, the models are not only optimally parsimonious, but also minimally redundant in the types of representations presupposed. There is no need for the parallel existence of schemes or rules and exemplars as in cognitive linguistics and similar approaches.

The class of data-oriented, fragment-based approaches (such as DOP and the fragment grammar approach of O'Donnell et al., 2011) is often compared to MBLP, but differs in two important aspects. Firstly, the DOP approach does not perform analogical reasoning, and can only manipulate and build on the fragments it has memorized. Secondly, its reliance on hierarchical structure even in the smallest fragments makes the approach fundamentally more complicated, also as a cognitive theory, than the memory-based approach that is open to structured representations, but does not presume them. In addition, an empirical comparison in the area of machine translation reported by Haque et al. (2011) shows that a fragment-based approach can be improved by analogical reasoning, by exploiting (rather than ignoring) the local context of source-side phrasal fragments, and finding proper nearest neighbours with likely translations, rather than finding all possible translations, including the unlikely ones.

As a counterbalance to claimed assumptions of the reliance of the human language processing system on structured representations or other abstract levels of representation and processing, the memory-based account has the advantage as a more parsimonious cognitive theory of

language processing. Yet, a large body of work in psycholinguistics has focused on testing notions of abstraction, and a considerable portion of this work has provided convincing empirical evidence for correlations of human behaviour with these abstract notions. Our view of this work is nuanced.

On the one hand, most work in MBLP and related approaches does embrace certain basic levels of abstraction. In our work we have mostly assumed letters and phonemes to be abstract working elements in human language processing, rather than segmented audio samples and letter images, in line with the work in psychology that has argued for these levels of abstraction in perception (e.g. for speech sounds Oden & Massaro, 1978; Toscano & McMurray, 2010; for visual letter reading Dehaene et al., 2001; McClelland, 1976) and its emergence in language learning and in learning to read. We also tend to use letter *n*-grams, words or word *n*-grams as abstract working units in our feature representations of examples.

On the other hand, we feel that much of the work that investigates word- and sentence-level processing and that advocates abstract representations and processing models (such as dual route models) ignores the fact that the functional behaviour of an example-based model may be very much like that of a more abstract model. Categorical decisions made by both models may often be identical, by virtue of the fact that the abstract model is intended to generalize over sets of examples – and thus may respond with the same prediction or decision as the analogical reasoning process would on the basis of the same set of nearest-neighbour examples the abstraction generalizes over. In other words, empirical evidence for an abstraction correlating with human behaviour may well be explained by an example-based model that happens to make the same predictions as the abstraction does, except that it makes them implicitly and temporarily rather than explicitly and permanently.

This explanation may cast a new light on the dual versus single route debate. One misunderstanding that permeates this debate is the confusion of memory-based routes based on table lookup or rote learning (given an input, one specific output is triggered), and analogical reasoning over exemplars, our MBLP approach. While the table lookup route is obviously incapable of overgeneralization or irregularization, the MBLP approach generates both; at the same time, it does produce table-lookup-like behaviour when encountering an exact match between a new example and a memorized example. The versatility of the MBLP approach makes it a viable candidate engine for single route models.

Acknowledgements

The authors wish to thank Steven Gillis, Gert Durieux, Maarten van Gompel and Peter Berck for their contributions to work summarized here, and to Royal Skousen, Harald Baayen, Colin Davis and Emmanuel Keuleers for valuable suggestions and discussions.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Notes

1. In all experiments referred to in the text, we used TiMBL (Tilburg Memory-Based Learner), a software package implementing several variants of memory-based classification, released under an open source license. TiMBL can be downloaded from http://ilk.uvt.nl/timbl.
2. There may be cases in which an exact match does not lead to proper analogical reasoning; this would be due to errors in either the training or the test data, or an inappropriate choice of features excluding some discriminatory feature that would yield the match non-exact.

References

Arnon, I., & Cohen, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

Beekhuizen, B., Bod, R., & Zuidema, J. (2013). Three design principles of language: The search for parsimony in redundancy. *Language and Speech*.

Bod, R. (1995). Enriching linguistics with statistics: Performance models of natural language. Ph.D. thesis, ILLC, Universiteit van Amsterdam.

Bod, R., Scha, R., & Sima'an, K. (Eds.). (2003). *Data-oriented parsing*. Stanford, CA: CSLI Publications.

Canisius, S., Van den Bosch, A., & Daelemans, W. (2006). Constraint satisfaction inference: Non-probabilistic global inference for sequence labelling. In *Proceedings of the EACL 2006 Workshop on Learning Structured Information in Natural Language Applications*, Trento, April 2006.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review*, *3*, 37–60.

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning* (pp. 115–123). San Mateo, CA: Morgan Kaufmann.

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, D. A. (2004). *Cognitive linguistics*. Cambridge, UK: Cambridge University Press.

Daelemans, W. (2002). A comparison of analogical modeling to memory-based language processing. In R. Skousen, D. Lonsdale, & D. B. Parkinson (Eds.), *Analogical modeling: An exemplar-based approach to language*. 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., Gillis, S., & Durieux, G. (2012). The acquisition of stress: A data-oriented approach. *Computational Linguistics*, *20*, 421–451.

Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Daelemans, W., & Van den Bosch, A. (2009). Memory-based learning. In A. Clark, C. Fox, & S. Lappin (Eds.), *Handbook of computational linguistics and natural language processing (pp*. 154–179*)*. Oxford, UK: Wiley-Blackwell Publishers.

Daelemans, W., Van den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine Learning*, *34*, 11–43.

De Saussure, F. (1916). Cours de linguistique générale, ed. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot.

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J. B., Le Bihan, D., & Cohen, L. (2001). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, *15*, 307–313.

Dresher, E., & Kaye, J. (1990). A computational learning model for metrical phonology. *Cognition*, *32*, 137–195.

Eddington, D. (2002). A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling. In R. Skousen, D. Lonsdale, & D. B. 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, NY: Oxford University Press.

Fix, E., & Hodges, J. L. (1951), *Discriminatory analysis—Nonparametric discrimination; consistency properties*, Technical Report Project 21–49–004, Report No. 4, USAF School of Aviation Medicine.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. New York, NY: Oxford University Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Haque, R., Naskar, S. K., Van den Bosch, A., & Way, A. (2011). Integrating source-language context into phrase-based statistical machine translation. *Machine Translation*, *25*, 239–285.

Hintzmann, D. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*, 400–401.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., & Herbst, E. (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.

Krott, A., Schreuder, R., & Baayen, R. H. (2002). Analogical hierarchy: exemplar-based modeling of linkers in Dutch noun-noun compounds. In R. Skousen, D. Lonsdale, & D. B. Parkinson (Eds.), *Analogical modeling: An exemplar-based approach to language*. Amsterdam, The Netherlands: John Benjamins.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning* (ICML-2001).

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* (pp. 487–495), Athens, Greece, 2009.

Lepage, Y., & Denoual, E. (2005). Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, *19*, 251–282.

Lepage, Y., & Shin-ichi, A. (1995). Saussurian analogy: A theoretical account and its application. In *Proceedings of the 16th International Conference on Computational Linguistics*, COLING-96 (pp. 717–722), Copenhagen, Denmark.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006) Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, *10*, 447–454.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

McClelland, J. L. (1976). Preliminary letter identification in the perception of words and nonwords. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 80–91.

McElree, B. (2006) Accessing recent events. In B.H. Ross (Ed.), *The psychology of learning and motivation* (Vol. *46*). San Diego, CA: Academic Press.

Mos, M., Van den Bosch, A., & Berck, P. (2012). The predictive value of word-level perplexity in human sentence processing: A case study on fixed adjective-preposition constructions in Dutch. In S. Gries & D. Divjak (Eds.), *Frequency effects in language learning and processing* (pp. 207–240). Berlin: De Gruyter Mouton.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *15*, 39–57.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

O'Donnell, T. J., Snedeker, J., Tenenbaum, J. B., & Goodman, N. D. (2011). Productivity and reuse in language. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, *33*, 1065–1076.

Paul, H. (1880). *Prinzipien der Sprachgeschichte* (1st ed.) Halle, Germany: Max Niemeyer.

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).

Punyakanok, V., Roth, D., Yih, W., & Zimak, D. (2005). Learning and inference over constrained output. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 1124–1129).

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. *14*, pp. 207–262). New York, NY: Academic Press.

Skousen, R. (1989). *Analogical modeling of language*. Dordrecht, The Netherlands: 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. E., & Medin, D. L. (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, UK: Cambridge University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434–464.

Van den Bosch, A. (1999). Careful abstraction from instance families in memory-based language learning. *Journal for Experimental and Theoretical Artificial Intelligence*, *3*, 339–368.

Van den Bosch, A., and Berck, P. (2009). Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical Linguistics*, *91*, 17–26.

Van den Bosch, A., & Daelemans, W. (2005). Improving sequence segmentation learning by predicting trigrams. In *Proceedings of the Ninth Conference on Natural Language Learning*, CoNLL-2005 (pp. 80–87), Ann Arbor, MI.

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.

Wiechmann, D. (2009). Understanding complex constructions: A quantitative corpus-linguistic approach to the processing of English relative clauses (Unpublished doctoral dissertation.) University of Jena, Germany.

Yvon, F., & Stroppa, N. (2007). Proportions in the lexicon: (Re) Discovering paradigms. *Lingue e linguaggio*, *2*, 201–226.

Zavrel, J., & Daelemans, W. (1997). Memory-based learning: Using similarity for smoothing. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 436-443). New Brunswick, NJ: Association for Computational Linguistics.

Zipf, G. K. (1935): *The psycho-biology of language: An introduction to dynamic philology*. Boston, MA: Houghton-Mifflin.