Towards a Psycholinguistic Computational Model for Morphological Parsing
Author(s): R. Harald Baayen and Robert Schreuder
Source: *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, Vol. 358, No. 1769, Computers, Language and Speech: Formal Theories and Statistical Data (Apr. 15, 2000), pp. 1281-1293
Published by: The Royal Society
Stable URL: http://www.jstor.org/stable/2666818
Accessed: 17/05/2011 12:17

THE ROYAL SOCIETY

Towards a psycholinguistic computational model for morphological parsing

BY R. HARALD BAAYEN AND ROBERT SCHREUDER

Interfaculty Research Unit for Language and Speech, Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands (baayen@mpi.nl; schreude@mpi.nl)

Psycholinguistic experiments on visual word recognition in Dutch and other languages show ubiquitous effects of word frequency for regular complex words. The present study presents a simulation experiment with a computational model for morphological segmentation that is designed on psycholinguistic principles. Results suggests that these principles, in combination with the presence of form and frequency information for complex words in the lexicon, protect the system against spurious segmentations and substantially enhance segmentation accuracy.

Keywords: morphological segmentation; full-form storage; mental lexicon

1. Introduction

When encountering the Dutch word *bestelauto*, Dutch readers understand this orthographic string to denote 'delivery van'. They hardly ever become aware of an alternative legitimate interpretation, 'berry counting car', corresponding to the segmentation *bes*+ *tel*+ *auto* instead of the correct segmentation *be*+ *stel*+ *auto*. Neither do readers seem to have any difficulty in discounting uninterpretable sequences of Dutch morphemes that likewise span the orthographic string, such as *bes* + *t* + *el* + *auto*. The question addressed in this study is how readers might accomplish the selection of the correct segmentation of a morphologically complex word.

The traditional approach in computational linguistics to morphological parsing proceeds in two steps. First, the set of possible segmentations that span the input string is calculated. Next, the combinatorial properties of morphemes are used to rule out illegal segmentations such as *bes* + *t* + *el* + *auto*, in which the verbal inflectional suffix *-t* follows a noun instead of a verb. In some statistically enhanced methods, co-occurrence frequencies are used to distinguish between probable parses ('delivery van') and improbable parses ('berry counting car').

The algorithm for determining the most probable segmentation described in the present paper is based on a rather different, psycholinguistically motivated conceptual framework, that of parallel lexical activation and lexical competition. The wetware of the human brain makes use of massively parallel and interactive processing, in contrast to the hardware of the present-day single-processor PC which operates sequentially.

The lexicon on which this algorithm operates also differs from the lexicons traditionally found in linguistics and computational linguistics. The traditional approach

in linguistics to the problem of morphological parsing is to assume that irregular complex words are stored in a lexicon along with the basic formative elements (stems and affixes), and that rules are used to segment regular complex words into their component constituents. However, various psycholinguistic studies report that high-frequency complex words are responded to more quickly and accurately in various experimental tasks than low-frequency words. This word-frequency effect has been obtained for both regular derived and regular inflected words in a range of languages (English, see Taft (1979), Sereno & Jongman (1997), Allegre & Gordon (1999); Dutch, see Baayen *et al.* (1997*b*, 2000*a*), Bertram *et al.* (2000), Schreuder *et al.* (1999); Italian, see Baayen *et al.* (1997*a*); and Finnish, see Bertram *et al.* (1999)). This word-frequency effect shows that human morphological processing is sensitive to the co-occurrence frequencies of constituents in regular complex words. The empirically observed knowledge of co-occurrence frequencies of morphemes in the mental lexicon is in line with statistically enhanced parsing models in computational linguistics that make use of conditional probabilities (hidden Markov models, Charniak (1993)) or databases with stored exemplars (data-oriented parsing, Bod (1998); lazy-learning based induction, Van den Bosch *et al.* (1996) and Daelemans *et al.* (1999)). The first aim of the present paper is to gauge the role that co-occurrence information may play in MATCHECK, a computational model for the identification of simplex and complex words, which is articulated within the psycholinguistic framework of parallel lexical activation and lexical competition (Baayen *et al.* 2000*b*; Baayen & Schreuder 1999). A second aim is to ascertain to what extent cognitive principles of human perception might contribute to enhancing segmentation accuracy.

In what follows, we first outline the basic mechanisms of this model. We then analyse the performance of the model by means of a simulation study using two different lexicons, one lexicon with only stems and affixes, and one lexicon with stems, affixes and, in addition, whole-word representations for regular complex words. We will show that segmentation performance is substantially enhanced when the latter lexicon is used in combination with an algorithm that implements the Gestalt principle that the whole has a perceptual advantage over its parts. We briefly review the differences between the kind of co-occurrence information that plays a role in our model and the kind of information used in Markovian statistical models. Finally, we discuss possible implications of our approach for natural language processing (NLP) tools.

2. A psycholinguistic segmentation model

MATCHECK distinguishes between a number of successive processing stages. The first stage concerns perceptual identification, the mapping of sensory visual or auditory information onto modality-specific form representations stored in long-term memory. We will refer to these form representations as *access representations*. Each access representation provides pointers to semantic and syntactic information in long-term memory.

During the second stage, the segmentation stage, access representations are activated over time by the sensory input. This segmentation stage is comparable with the stage of lexical look-up by means of, for example, an L-tree (Sproat 1992). Once an access representation reaches a given threshold activation level, it is copied to a short-term memory buffer.

Once the representations in the short-term memory buffer provide full, non-overlapping spannings of the input, these segmentations are passed on to the following processing stages of licensing (the checking of subcategorization compatibilities), composition (the compositional computation of the meaning of the whole from its parts) and semantic activation (the co-activation of semantically related representations in long-term memory).

At present, the only processing stage that is computationally implemented is the segmentation stage, and it is the performance of this segmentation stage and the selection of the most appropriate segmentation of the input that is the topic of the present paper.

The segmentation model implements the activation metaphor that is commonly used in psycholinguistic modelling. This metaphor is meant to capture the experimental observation that, in human processing, information comes in over time, and that its availability is not an all-or-nothing question, but rather that of accumulating evidence. In the activation framework, access representations are assigned resting activation levels that are proportional to their frequency of occurrence. Thus, high-frequency words will reach a preset activation threshold more quickly than low-frequency words.

One way in which we can gauge the behaviour of the segmentation model is to study the different time-steps in the model at which full spannings of the input become available. It is hoped that the order in which such full spannings become available over time reflects their ranking in terms of correctness and semantic plausibility. Thus far, this issue has not been a topic of systematic investigation. Before addressing this issue by means of a simulation study in § 3, however, we first introduce some more technical details.

In MATCHECK, affixes, stems and full forms have modality-specific access representations with activation weights $a(w,t)$. Lexical competition is modelled by imposing a probability measure on the activation weights of the access representations for a given time-step t. Once the probability of identification $p_{w,t}$ of an access representation w,

$$p_{w,t} = \frac{a(w,t)}{\sum_{i=1}^{V} a(w_i,t)}, \tag{2.1}$$

with V the number of access representations in the lexicon, exceeds a pre-set probability threshold θ, it is copied into the short-term memory buffer.

With immutable activation weights, we would have a completely static system. Activation weights change, however, in two ways. Once an access representation has reached threshold activation level, its activation weight will begin to decrease, thereby effectively freeing activation probability for other access representations.

The second way in which the system is made dynamic is by having access representations enter activation decay at a moment in time that is proportional to their similarity to the target word. The similarity of an access representation to the target word in the input is defined by means of a similarity metric based on an edit distance measure. Words that are very similar to a given target word will have increasing activation weights for longer numbers of time-steps than words that are very dissimilar. Consequently, the onset of activation decay is located earlier in time for dissimilar words than for similar words. A detailed initial formal definition of MATCHECK is

available in Baayen *et al.* (2000*b*). In what follows, we briefly present the main concepts and the way in which they have been refined in the current version of the model.

The similarity metric defines a span of time-steps during which an access representation is 'on hold', i.e. during which its activation weight is allowed to increase. Let the indicator function $\mathcal{H}(w,t)$ be 1 when the access representation w is 'on hold' at time-step t, and 0 otherwise. We express the activation weight of w at t as a function of its activation weight at the previous time-step. Starting with an initial activation weight equal to its frequency f_w in a given corpus, $a(w,0) = f_w$, we have

$$a(w,t) = I_{[\mathcal{H}(w,t)=1]}\frac{a(w,t-1)}{\delta_{w_i}} + I_{[\mathcal{H}(w,t)=0]}[a(w,0) + \delta_{w_i}\{a(w,t-1) - a(w,0)\}]. \tag{2.2}$$

The first term in (2.2) specifies that if an access representation is on hold, its new activation weight is the product of its activation weight at the previous time-step and the reciprocal of its decay rate, δ_{w_i}. The second term in (2.2) implements asymptotic decay to the original resting activation level at the initial time-step $t = 0$ for access representations that are no longer on hold.

Each access representation w_i is assigned its own decay rate δ_{w_i} and activation rate $\delta_{w_i}^{-1}$, which specifies how quickly words become activated and how quickly they decay again. The value of the by-item decay rate, δ_{w_i}, is determined by two opposing principles. The first principle assigns higher decay rates to shorter and more-frequent morphemes, enforcing rapid activation and rapid decay. This principle is implemented by means of the function $g(\delta,\alpha)$. Let $L(w_i)$ denote the length of w_i in letters, let f_{w_i} denote the frequency of w_i, and let δ be the baseline decay rate. We can now define $g(\delta,\alpha)$ as follows:

$$g(\delta,\alpha) = \delta\frac{L(w_i)}{L(w_i) + (\alpha/L(w_i))\log(f_{w_i})}. \tag{2.3}$$

Large values of the 'Spike' parameter α, $\alpha > 0$, lead to lower decay rates and higher activation rates, especially so for longer words, as exemplified in the right-hand panel of figure 1, leading to spike-like activation patterns over time. By means of α, we can enhance the identification of short high-frequency inflectional affixes.

The second principle that co-determines the decay rate of access representations implements the Gestalt principle that the whole takes precedence over the parts in recognition: 'forest before trees' (Navon 1977). This principle is realized by the 'forest' parameter ζ, $\zeta > 0$, in the function $f(\delta,\zeta)$. Denoting the target word by T, we have:

$$f(\delta,\zeta) = \begin{cases} \delta + (1-\delta)\left(\dfrac{|L(w_i) - L(T)|}{\max(L(w_i),L(T))}\right)^{\zeta}, & \text{iff } \zeta > 0, \\ \delta_i, & \text{otherwise.} \end{cases} \tag{2.4}$$

The left-hand panel of figure 1 illustrates, firstly, that words that are much shorter or much longer than the target word receive smaller activation rates, and, secondly, that decreasing the value of ζ leads tc an increased contrast in activation rate between words similar and dissimilar in length to the target word. Smaller values of ζ lead

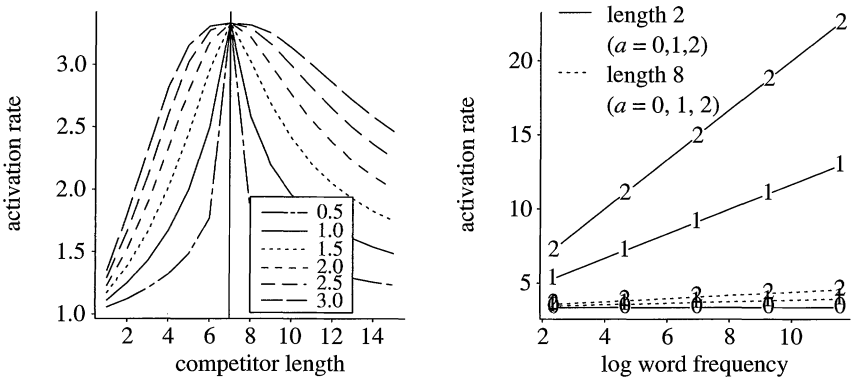

Figure 1. The effect on the activation rate ($1/\delta$) of the forest–trees parameter ζ (left panel, $\zeta = 1.0, 1.5, 2.0, 2.5, 3.0$, for a target word with length 7, $\delta = 0.3$, and competitor access representations with lengths ranging from 1 to 15) and the Spike parameter α (for word lengths 2 and 8 with as baseline $\delta = 0.3$).

to a bigger advantage of the whole over its parts. Finally, we now define the by-item decay rate, δ_{w_i}, by composition of the two functions for the two principles:

$$\delta_{w_i} = f(g(\delta, \alpha), \zeta). \tag{2.5}$$

As shall become clear below, ζ and α are crucial in order to allow the model to make optimal use of co-occurrence information simultaneously with efficient segmentation.

The definitions of $a(w,t)$ and the decay rate δ_{w_i} differ from those given in Baayen *et al.* (2000*b*). In their initial definition of MATCHECK, there is only the general decay rate for all words, irrespective of their frequency and length. The revised definition of how the activation weights change, equation (2.2), is a first step towards modelling activation weight as a function of frequency of exposure. This necessitated modelling lexical competition by means of raising the activation weights of compatible access representations instead of decreasing the activation weights of incompatible access representations, as in Baayen *et al.* (2000*b*).

An important psycholinguistic feature of the model is the prominence assigned to the left and right edges of words (see Cutler *et al.* (1985) for a review of experimental evidence). The activation weights of access representations are increased only for constituents that are aligned with the left or right edge of the word, or that are aligned with access representations that have reached threshold and that themselves are edge aligned, either with the word edge itself or with another edge-aligned constituent in the short-term memory buffer. In this way we avoid creating a system in which *rest* in *prestige* becomes fully activated, because *ige* is not a legitimate constituent. Conversely, *bestel* ('deliver') in ***bestel****auto* ('delivery van') quickly becomes available for activation increase through a part–whole Gestalt principle once either *auto* or *bestel* have reached activation threshold.

Figure 2 illustrates the time course of the probabilities of identification for selected access representations when *bestelauto*, 'delivery van', is presented to the model in the visual modality. The first access representation to reach threshold is the high-frequency prefix *be-*, followed by *auto*, 'car', the verb *bestel*, 'deliver', the full form *bestelauto*, and various other embedded words such as *best*, 'best', and *bes*, 'berry'.

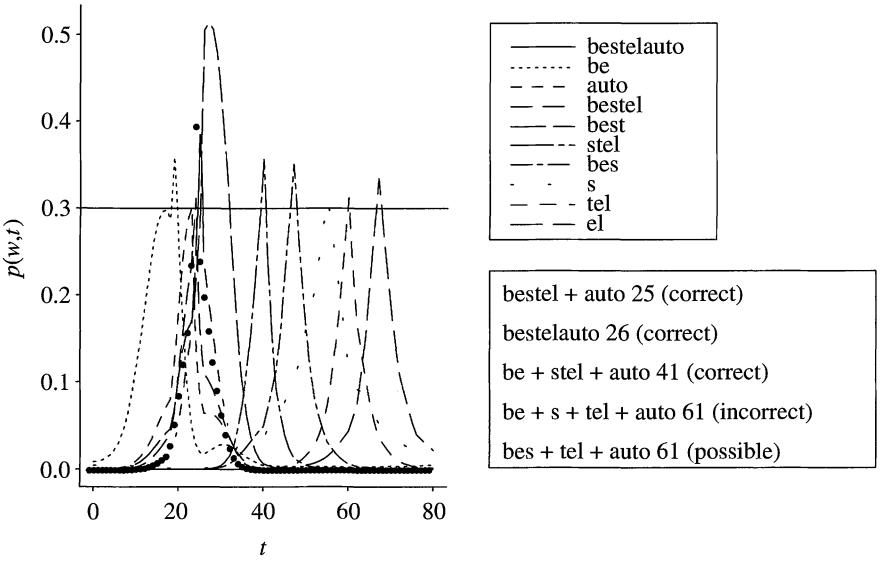

Figure 2. Probability of identification $p(w,t)$ for selected access representations as a function of time-step t, with activation threshold $\theta = 0.3$, for *bestelauto*, 'delivery van'. The time-steps at which full spannings become available are listed in the lower right-hand corner.

The first full spannings become available at time-step 25, *bestel* + *auto*, at time-step 26, *bestelauto*, and *be*+*stel*+*auto* at time-step 41. These are all correct analyses. Incorrect segmentations follow 20 or more time-steps later.

Note that, although full-form knowledge is available in the form of access representations for *bestelauto* itself, the first analysis to become available is based on a segmentation of the input into its immediate constituents and not the full form. However, depending on the complexity of subsequent lexical processing at the licensing and composition stages, it may happen that it is nevertheless a full-form representation that is ultimately the first to activate the full semantics of the target word. In this sense, our model is a dual route model in which an access route based on full-form information runs in parallel with an access route based on decomposition. We are currently implementing the post-segmentation processes of licensing and composition in an explicit computational model.

3. Segmentation performance

How well does the segmentation module of MATCHECK succeed in selecting probable segmentations and in assigning a low priority to improbable and incorrect segmentations? To gauge the performance of the model, we randomly selected two sets of complex words from the CELEX lexical database. The first set contained 200 words with an orthographic length in the range of 5–12 letters. The second set contained 100 words with an orthographic length in the range 13–20. For both sets, we compared the performance of the model using two lexicons: a full-form lexicon with 98 430 entries, and a parse-only lexicon with 15 015 entries. The parse-only lexicon is a subset of the full-form lexicon that only contains simplex words and affixes. The full-form lexicon contains entries for forms such as *bestel* and *bestelauto* in addition

Table 1. *Statistics for the kinds of segmentations for the two test sets*

(Mean, median and range of the number of correct, possible, and incorrect segmentations by words. Count: the number of words with at least one correct/possible/incorrect segmentation.)

Enhanced race model

	word lengths 5–12			word lengths 13–20		
	correct	possible	incorrect	correct	possible	incorrect
mean	3.1	0.3	3.3	3.2	0.7	8.7
median	3.0	0.0	2.0	3.0	0.0	5.5
rangeL	1	0	0	1	0	0
rangeU	8	3	24	10	6	69
count	200	44	151	100	30	85

Full parsing model

	word lengths 5–12			word lengths 13–20		
	correct	possible	incorrect	correct	possible	incorrect
mean	1.2	0.2	2.4	1.4	0.3	7.9
median	1.0	0.0	1.0	1.0	0.0	4.0
rangeL	0	0	0	0	0	0
rangeU	4	2	19	6	4	51
count	197	35	146	96	16	82

to the basic morphemes *be-*, *stel*, and *auto*. The full-form lexicon contains full-form representations up to and including a word length of 12 letters, in order to gauge the performance of MATCHECK for novel forms for which it cannot rely on stored information in its lexicon. Both lexicons were derived from the CELEX lexical database (Baayen *et al.* 1995), which is based on a corpus of 42 million words of written Dutch. The initial activation weights of the entries in the lexicons are identical to the frequencies of the corresponding words in this corpus. Affixes were assigned initial activation weights equal to the summed frequencies of complex words in which they appear as constituents in CELEX. Allomorphy is handled by separate listing of allomorphic variants. Thus, using an English example, a form such as *easier* is assigned two representations in the full-form lexicon, the full-form representation (*easier*), and, as a representation of the base, the orthographic allomorph *easi*.

For the set of shorter words, the average number of words and morphemes in the full-form lexicon embedded in these words was 12.8 (range 4–33). For the set of longer words, the average number of embedded morphemes and words was 21.9 (range 10–33).

Table 1 summarizes the main performance characteristics for the two test sets, for two different simulation experiments. The first simulation experiment uses the full-form lexicon and the revised model definition outlined in the previous section with $\delta = 0.3$, $\alpha = 1.2$, $\zeta = 1.5$ and $\theta = 0.3$. We will refer to this simulation as the 'enhanced

Table 2. *Segmentations available at first time-step*

(The first segmentations to become available for the two test sets of words (lengths 5–12 and lengths 13–20) using a full-form lexicon and a parsing lexicon.)

	enhanced race		full parsing	
word length	5–12	13–20	5–12	13–20
correct segmentation(s) only	194	82	133	55
possible segmentation(s) only	5	3	5	1
incorrect segmentation(s) only	1	4	30	12
combinations of segmentation	0	11	32	32
total number of segmentations	200	100	200	100

race model'. The second simulation experiment uses the full parsing lexicon, the smaller lexicon that only contains simplex words and affixes. Here, we disabled the spike and forest options by setting α and ζ to zero, in order to ascertain the behaviour of the model in its simplest form. We will refer to this experiment as the 'full parsing model'. The rows labelled 'mean' and 'median' present the mean and median number of correct, possible and incorrect segmentations. For our working example, *bestelauto*, examples of correct analyses are *bestelauto* and *bestel+auto*, an example of a possible but implausible segmentation is *bes + tel + auto*, and an example of an incorrect segmentation is *bes+t+el+auto*. Table 1 also lists the corresponding lower ('rangeL') and upper ('rangeU') ranges. Finally, the rows labelled 'count' present the counts of words for which at least one correct segmentation was generated, the counts for which at least one possible segmentation was generated, and the counts for which at least one incorrect segmentation was produced. The 'count' row for the full parsing model shows that the model failed to produce a correct parse for $200 - 197 = 3$ words of length 5–12.

Note that, unsurprisingly, the longer words generally have larger numbers of segmentations, especially so in the case of Incorrect segmentations. Also note that the numbers for the enhanced race model are slightly larger than those for the full parsing model, not only for the correct and possible segmentations, but also for the incorrect segmentations. Apparently, the larger number of words in the full-form lexicon, among which we find morphologically correct substrings such as *bestel* ('deliver') in *bestelauto* ('delivery van'), does not, by itself, lead to an *a priori* numerical advantage for correct segmentations. Finally, note that the summed numbers of possible and incorrect segmentations are larger than the numbers of correct segmentations in the mean and, except for word length 5–12 in the enhanced race model, also in the median. Generalizing over the two lexicons in the two experiments, we may conclude that the probability of selecting a correct segmentation at random is less than 0.5.

How well does MATCHECK succeed in selecting correct parses? The way in which MATCHECK assigns a ranking to different segmentations is in the order in which it makes the segmentations available over time. Ideally, the first segmentation to become available should be the correct one. The later a segmentation arrives, the less likely it is to be useful for further (human) processing.

Table 2 classifies the kinds of segmentations that are the first to become available. First, consider the full parsing model. For this simulation, 133 of the 200 words of length 5–12 are assigned one or more correct segmentations, without any other kinds of segmentations becoming available at the same time-step. The corresponding numbers for possible (but implausible) and incorrect segmentations are 5 and 30, respectively. Finally, 32 of the 200 words show combinations of correct and incorrect segmentations. For the 100 words of length 13–20, performance drops, with, for example, an increase from 16% to 32% of 'ambiguous' time-steps with multiple kinds of segmentations, and with only 55% instead of 67% of the words being assigned correct segmentation(s) exclusively.

Turning to the enhanced race model, we observe a much higher success rate. For words of length 5–12, 194 out of 200 emerge with exclusively correct segmentations. Of the remaining five words, moreover, four are assigned implausible but linguistically legal segmentations. Ambiguities due to the simultaneous presence of different kinds of segmentations, both correct and incorrect, do not arise. Interestingly, less than half of the correct segmentations (92 out of 200) can be attributed to full forms being the first to become available. This shows that our high success rate is not due to the 'trivial' full-form segmentation being always the first to arrive.

The performance of MATCHECK for the 100 longer words is less accurate. This is probably due to two factors: the larger numbers of possible segmentations for longer words (see table 3); and the absence of full forms in MATCHECK's lexicon. Even though these words are well-established words according to CELEX, we have given them the status of neologisms in the simulation in order to gauge how well MATCHECK performs on unseen words. For 82 out of 100 words, a correct segmentation is the first to arrive, a considerable improvement over the 55 out of 100 for the full parsing model. In the 11 cases in which combinations of segmentations arrive, a correct segmentation is always present. And if we allow ourselves to exclude as *a priori* incorrect those segmentations in which exclusively word final morphemes appear in word initial position in a segmentation, the number of correct segmentations increases to 94 out of 100. This suggests that further enhancements of the segmentation module are worth developing, especially as the segmentation stage of MATCHECK does not make use of any linguistic information at all, either semantic information to rule out possible but improbable segmentations, nor subcategorization information to rule out incorrect segmentations.

4. Comparison with statistical language models

How does the present psycholinguistically motivated segmentation algorithm compare with standard statistical approaches using Markov models, for which only weak psycholinguistic claims are made? The two approaches share the belief that co-occurrence probabilities are part and parcel of morphological parsing. The two approaches differ in two respects. First, the accuracy of MATCHECK crucially depends on cognitive insights such as the 'forest-before-trees' Gestalt principle. Second, the two techniques differ with respect to what kind of co-occurrence probabilities are used. As MATCHECK assigns segmentations a ranking in terms of the time-step at which a full spanning becomes available without specification of the internal hierarchical structure of the segmentation, we compare its temporal probability ranking with the probabilities assigned to strings of morphemes by a simple hidden Markov

Table 3. *Morphological n-gram frequencies and probabilities*

(Frequencies and probabilities of the morphological unigrams, bigrams, and trigrams in *onwerkbaar*, 'unworkable'.)

	frequencies	probabilities
unigram		
on-	125 564	$\Pr(on\text{-}) = 125\,564/N$
werk	21 018	$\Pr(werk) = 21\,018/N$
-baar	37 721	$\Pr(\text{-}baar) = 37\,721/N$
bigram		
onwerk	256	$\Pr(werk \mid on\text{-}) = 256/125\,564 = 0.0020$
werkbaar	34	$\Pr(baar \mid werk) = 34/21\,018 = 0.0016$
trigram		
onwerkbaar	9	$\Pr(baar \mid on\text{-}, werk) = 9/256 = 0.0352$

model using trigrams and smoothing with unigrams and bigrams (Charniak 1993, p. 40),

$$\Pr(w_n \mid w_{n-2,n-1}) = \lambda_1 \Pr(w_n) + \lambda_2 \Pr(w \mid w_{n-1}) + \lambda_3 \Pr(w \mid n-2, n-1), \quad (4.1)$$

with $\sum_i \lambda_i = 1$. Using as an example the complex word *onwerkbaar*, 'unworkable', we calculate the morpheme unigram, bigram and trigram probabilities from the corresponding frequencies in a given corpus of size N, as shown in table 4.

The bigram frequency for *on+werk* is obtained by summation of the frequencies of *onwerkbaar* itself, *onwerkelijk*, 'unreal', and *onwerkzaam*, 'ineffective'. (The sequence *on+werk* does not appear independently in Dutch.) Similarly, the bigram frequency for *werkbaar* is obtained by summation of the frequencies of *werkbaar*, *onwerkbaar* and *verwerkbaar*, 'processable'. The likelihood of the sequence *on+werk+baar* in a hidden Markov model equals the product of $\lambda_1 \Pr(on\text{-})$, $\lambda_1 \Pr(werk) + \lambda_2 \Pr(werk \mid on\text{-})$ and $\lambda_1 \Pr(\text{-}baar) + \lambda_2 \Pr(\text{-}baar \mid werk) + \lambda_3 \Pr(\text{-}baar \mid on\text{-}, werk)$, using the probabilities in table 3 and assuming that estimates for the smoothing parameters, λ_i, are available.

These probabilities differ from those that are allowed to play a role in MATCHECK as initial probabilities of identification.

Thus, first, the sequence *on+werk*, which has a non-zero bigram probability in the hidden Markov model, is not represented by an independent access representation in our psycholinguistic model, because *onwerk* is not an existing word in Dutch.

Second, even though the probability $\Pr(\text{-}baar \mid werk)$ is paralleled by an access representation for *werkbaar*, the access representation of *werkbaar* receives, as its initial activation weight, the frequency of *werkbaar*, and not the summed frequencies of *werkbaar*, *onwerkbaar*, and *verwerkbaar*. The reason for not cumulating the frequencies of *onwerkbaar* and *verwerkbaar* with the frequency of *werkbaar* is that recent experimental studies of cumulative token frequency effects have revealed that a type count, but not a token count, of morphological descendants co-determines response latencies, as shown by Schreuder & Baayen (1997) for simplex words and De Jong *et al.* (2000) for complex words.

Third, the probability that *on-* is followed by *werkbaar*, Pr(*werk, baar* | *on-*), does not play a role in the hidden Markov model, while in MATCHECK the access representations of *on-* and *werkbaar* are free to combine to deliver a full spanning for *onwerkbaar*.

Thus, our present approach is more similar to statistical methods that make use of lazy learning (Van den Bosch *et al.* 1996; Daelemans *et al.* 1999) or data-oriented parsing (Bod 1998) than to techniques based on hidden Markov models, in that co-occurrence sequences are only taken into account when they represent attested constituents.

5. Discussion

When coupled with a full-form lexicon, MATCHECK reveals good performance (97% correct initial segmentations, of which slightly less than half are due to full-form segmentations) for known words and reasonable performance (82–94% correct initial segmentations) for neologisms. Focusing on the words with length 5–8, for which full forms are available in the lexicon, we observe a reliable correlation of the time-step at which a correct segmentation becomes available and (log) word frequency ($r = -0.51$, $t(196) = -8.37$, $p < 0.0000$), mirroring the correlations between response latencies and word frequency in various psycholinguistic experimental tasks (see, for example, Bertram *et al.* 2000*b*). Although the segmentation performance of MATCHECK is surprisingly good, it will probably prove impossible to enhance the performance to a full 100% for any kind of complex words. The model needs to be enriched with additional modules that exploit subcategorization and semantic knowledge in order to handle adequately those possible and incorrect segmentations that happen to arrive before, or simultaneously with, correct segmentations.

To what extent might MATCHECK be useful for general NLP purposes, as opposed to psycholinguistic modelling? Given the high degree of accuracy with which MATCHECK assigns priority to the correct segmentations, we expect that our algorithm might be useful in NLP involving morphological processing, as, for instance, in text-to-speech systems. Possibly, the implementation of well-motivated psychological cognitive principles in language engineering may lead to improved tools, just as the incorporation of standard statistical techniques leads to enhanced performance.

However, running MATCHECK on large full-form lexicons is very time consuming. On a Sun Sparc Ultra-10 elite 30 workstation, the segmentation of a single complex word requires roughly 1.5 min. Interestingly, we have found segmentation performance to be nearly equally accurate when the algorithm is applied not to all 98 430 words in our full-form lexicon, which we did for psycholinguistic reasons, but just to the set of embedded strings (including the full form, if present) in the lexicon for a given test word. In this case the program takes only a few seconds to complete. This suggests that it might be feasible to incorporate the MATCHECK algorithm in morphological parsers used in practical NLP tools.

Whether the MATCHECK algorithm will also be useful outside the domain of morphological segmentation in, for instance, sentential parsing, is unclear. The psycholinguistic principles that we have built into MATCHECK and that have considerably improved its performance are specific to the domain of lexical processing, the domain in which, at least in human language processing, the role of storage of form and meaning for complex linguistic structures is most prominent. What we have

learned from the present study is that, paradoxically, it is precisely this storage of full-form information in the lexicon that enhances morphological segmentation.

This research was supported by a Pionier grant of the Dutch National Research Council (NWO) to the first author.

References

Allegre, M. & Gordon, P. 1999 Frequency effects and the representational status of regular inflections. *J. Memory Lang.* **40**, 41–61.

Baayen, R. H. & Schreuder, R. 1999 War and peace: morphemes and full forms in a non-interactive activation parallel dual route model. *Brain Lang.* **68**, 27–32.

Baayen, R. H., Piepenbrock, R. & Gullikers, L. 1995 *The CELEX lexical database.* (CD-ROM) Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., Burani, C. & Schreuder, R. 1997*a* Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. In *Yearbook of Morphology 1996* (ed. G. E. Booij & J. van Marle), pp. 13–34. Dordrecht: Kluwer.

Baayen, R. H., Dijkstra, T. & Schreuder, R. 1997*b* Singulars and plurals in Dutch: evidence for a parallel dual route model. *J. Memory Lang.* **36**, 94–117.

Baayen, R. H., Schreuder, R., De Jong, N. & Krott, A. 2000*a* Dutch inflection: the rules that prove the exception. In *Storage and computation in the language faculty* (ed. S. Nooteboom, F. Weerman & F. Wÿnen). Dordrecht: Kluwer. (In the press.)

Baayen, R. H., Schreuder, R. & Sproat, R. 2000*b* Modeling morphological segmentation in a parallel dual route framework for visual word recognition. In *Lexicon development for speech and language processing* (ed. F. van Eynde & D. Gibbon). Dordrecht: Kluwer.

Bertram, R., Laine, M., Baayen, R., Schreuder, R. & Hyönä, J. 1999 Affixal homonymy triggers full-form storage even with inflected words, even in a morphologically rich language. *Cognition* **74**, B13–B25.

Bertram, R., Schreuder, R. & Baayen, R. 2000 The balance of storage and computation in morphological processing: the role of word formation type, affixal homonymy, and productivity, *J. Exp. Psych. Memory Learning Cognition* **26**, 1–23.

Bod, R. 1998 *Beyond grammar: an experience-based theory of language.* Stanford, CA: CSLI.

Charniak, E. 1993 *Statistical language processing.* Cambridge, MA: MIT Press.

Cutler, A., Hawkins, J. A. & Gilligan, G. 1985 The suffixing preference: a processing explanation. *Linguistics* **23**, 723–758.

Daelemans, W., Zavrel, J., Van der Sloot, K. & Van den Bosch, A. 1999 TiMBL: Tilburg memory based learner reference guide 2.0. Computational Linguistics Tilburg University report 99-01.

De Jong, N., Schreuder, R. & Baayen, R. 2000 Family size effects for uninflected and inflected verbs. *Language and Cognitive Processes.* (In the press.)

Navon, D. 1977 Forest before trees: the precedence of global features in visual perception. *Cognitive Psych.* **9**, 353–383.

Schreuder, R. & Baayen, R. H. 1997 How complex simplex words can be. *J. Memory Lang.* **37**, 118–139.

Schreuder, R., De Jong, N., Krott, A. & Baayen, R. 1999 Rules and rote: beyond the linguistic either-or fallacy. *Behavioral Brain Sci.* **22**, 1038–1039.

Sereno, J. & Jongman, A. 1997 Processing of English inflectional morphology. *Memory Cognition* **25**, 425–437.

Sproat, R. 1992 *Morphology and computation.* Cambridge, MA: MIT Press.

Taft, M. 1979 Recognition of affixed words and the word frequency effect. *Memory Cognition* **7**, 263–272.

Van den Bosch, A., Daelemans, W. & Weijters, T. 1996 Morphological analysis as classification: an inductive learning approach. In *Proc. NEMLAP 1996, Ankara*, pp. 59–72.

Discussion

N. Ostler (*Linguacubun Ltd, Bath, UK*). You discussed English and Dutch morphology, but will your approach also apply to highly inflected agglutinative languages?

R. H. Baayen. We are collaborating with colleagues in Finland to explore this question. Some results have been obtained on shorter words. My Finnish colleagues think that, for such shorter words, the only thing that takes place in visual processing is the storage of derived forms and compounds, but not inflections.

F. Pereira (*AT & T Laboratories, Florham Park, NJ, USA*). Is your contrast of the Bloomfieldian view with your own tabulating one not somewhat excessive?

R. H. Baayen. Well, perhaps, but proponents of the Bloomfieldian view find our data very problematic in the sense that they contradict everything that they believe to be true about how the human cognitive system works, namely that there are symbolic rules that do the actual processing with no storage mechanism.

R. Rosenfeld (*Carnegie Mellon University, Pittsburgh, PA, USA*). I wonder to what extent your inspiration from the architecture of the human brain had any effect on the success of the experiments that you described. Are there other competing statistical methods that are not inspired by the connectionist way of thinking? Are you aware of any model in the domain of psycholinguistics that has really addressed the issue of computational tractability?

R. H. Baayen. Connectionist models, which are excellent pattern matchers, do not work with current technology when large lexicons (greater than 5000 words or so) are used. However, the purpose of such experiments in psycholinguistics is not to test if a model will work in the real world, but to see if it explains what one sees in the lab. What we are trying to do is to see if one can build a model that can do a reasonable job in actual segmentation work, and at the same time use an algorithm that is psycholinguistically motivated.

K. I. B. Spärck Jones (*University of Cambridge, UK*). Could you say something about where the frequency data is obtained from and how much data one needs?

R. H. Baayen. We used around 40 million words as source data. Since the source comes from the written language, one might get better evidence if spoken language data were used instead.

M. Huckvale (*University College London, UK*). Could you comment on the role of full forms and why performance is improved in their presence?

R. H. Baayen. Full forms are merely present; hence, the amount of probability in the system for the right kind of form is increased. The full forms protect the system against many possible, but incorrect, parses containing high-frequency morphs. But the system retains the capability to analyse neologisms, not just full forms.