Frequency in morphology

András Kornai

In: I. Kenesei (ed): Approaches to Hungarian Vol 4 (1992) 246-268

0 Introduction

The recent work in statistical parsing (Church 1988, Schabes 1991) and statistical machine translation (Brown *et al* 1990) calls the traditional rule-based view of grammar into question. These authors emphasize that grammatical rule systems aiming at syntax-directed translation, and even rule systems aimed at the description of a single language, break down when faced with the actual complexity of natural language data. In fact, under realistic testing conditions the "example-based" or "corpus-based" systems that employ some general-purpose optimization algorithm in order to extract statistical regularities from the data fare just as well as the rule-based systems in which the regularities are extracted beforehand by the grammarian. In the light of these facts it is natural to extend the inquiry to morphology and ask how statistical morphological systems that exploit the frequency information in the data will compare with rule-based morphological systems that exploit the expertise of the grammarian.

The paper reports the results of a pilot study performed on the largest extant machine-readable corpus of a morphologically complex language, namely the SZO1TA1R corpus (see Kornai 1986) based on the Debrecen Thesaurus (Papp 1969) and the Frequency Dictionary of Hungarian (Füredi and Kelemen 1989). Section 1 presents the necessary background information about statistical approaches to grammar in general, about the traditional position-class view of morphology and its modern generalizations (Kiparsky 1982, Koskenniemi 1983), and about the Hungarian nominal and verbal paradigms (Antal 1961, 1966). Section 2 presents a simple statistical method to estimate the number of inflected word tokens required for "saturated paradigms" i.e. stems for which all paradigmatic forms are actually attested in the corpus. Although the SZO1TA1R corpus is to small to contain fully saturated paradigms, using an empirical finding of the pilot study, namely that linguistically orthogonal morphosyntactic features are statistically independent, we can still estimate the required corpus size. In order to saturate the paradigms of the most frequent 1,000 verb stems we need to collect a corpus 630 to 800 times bigger than the present one (which is .5m words), and in order to saturate the paradigms of the 30 most frequent nouns we would need over 33 billion words.

Do we really need all the paradigmatic forms to be exemplified in our corpus to be able to build them into our statistical morphology? The concluding Section 3 proposes hybrid Hidden Markov Model (HMM) systems that bring both grammatical knowledge and statistical methods to bear on the problem of acquiring morphological regularities from sparse data, and argues that such systems can avoid excessive reliance on the expertise of the grammarian and at the same time require more modest corpora than the "brute force" approach assumed earlier.

1 The theoretical background

Theoretical linguistics traditionally pays very little attention to frequency data: the fact that certain forms appear considerably more often than other equally grammatical (or equally ungrammatical) forms is seldom mentioned, let alone explored, outside of studies pertaining to dialectal/sociolectal

variation, to child language acquisition, or to psycholinguistics. The mainstream generative view takes grammars to be algebraic rule systems that manipulate (rewrite) or directly characterize (constrain) structures built from discrete symbols.

This view can be contrasted to the *statistical* view that takes grammars to be statistical regularities obtaining between ensembles of continuous variables. For the moment we will leave unresolved the larger issue whether such regularities are part of linguistic competence (the position taken in Labov (1965) and subsequent variationist work) or fall into the domain of performance (as assumed in Chomsky (1965) and subsequent generativist work) and concentrate on a more mundane question: where does the grammar come from? In the generative tradition, the rules, constrains and representations constituting the grammar are devised by the grammarian. In the statistical tradition, the regularities that make up the grammar are not devised by the grammarian: rather, they are found by general-purpose statistical optimization methods, with little or no human intervention.

It is instructive to see how these methods fare in various domains (modules) of grammar. In phonetics, and in particular in speech recognition, statistical models are clearly superior to "expertise-based" models (Makhoul and Schwartz 1986). In speech synthesis, models based on rules devised by the grammarian are slightly better than statistical models, but in other phonetic tasks such as speaker identification or voice compression again the best models are statistical in nature. In phonology and morphology expertise-based models predominate, but statistical (in particular, neural net) models are becoming increasingly more important (Legendre *et al* 1990, Goldsmith 1992, Prince 1992, MacWhinney, this volume).

In syntax the theoretical research is still largely conducted in the rule/constraint paradigm, but the applied syntactic work no longer relies on the results of this theoretical work. In fact, it is now a commonplace in Computational Linguistics that generative grammar based parsers do not work. The tradition of statistical Finite State Grammars (Markov 1913) has therefore been extended to Context Free Grammars (see e.g. Lari and Yound 1990), and Tree Adjoining Grammars (Schabes 1991) and it is fair to say that the field as a whole does not view the grammarian as a reliable source of syntactic information.[1] Finally semantics, long the stronghold of the rule-based "hard AI" approach, has seen the first serious statistical challenges in applications such as word sense disambiguation and machine translation (Brown *et al* 1990).

While the best way to deal with the higher modules is by no means settled, the undeniable success of the statistical approach at the lowest (phonetic) module makes the next lowest (lexical) module a natural candidate for extending the domain of statistical language models. In Section 2 we will see why the simplest "brute force" statistical approach to phonology/morphology is doomed to failure, and in Section 3 an HMM approach with a greater chance of success will be outlined. The remainder of Section 1 is devoted to a brief overview of the standard approaches and of the basic facts of Hungarian: readers familiar with linguistic theory and/or with Hungarian phonology/morphology can skip 1.1 and/or 1.2 without great loss.

1.1 Three models of morphology

Generative linguistics initially denied the existence of an autonomous morphological component or module of the grammar (Chomsky 1957, 1965, Chomsky and Halle 1968) and the contemporary sense of the "lexicon" as a joint repository of phonological and morphological rules, constraints, and representations did not fully emerge until the advent of Lexical Phonology and Morphology (Kiparsky 1982). Thus it is not surprising that the dominant computational morphological model of the period, Koskenniemi's two-level Phonology (1983) was inspired not so much by the (re)emerging generative theory of morphology as by the more traditional structuralist model of *position class morphology*.

[1]The idea that the expertise of the linguist is not particularly relevant to practical tasks that are linguistic in nature does not seem to be as farfetched as even ten years ago. For example, it seems quite likely that in another ten years the expertise of grandmasters will no longer be considered relevant in winning a game of chess.

1.1.1 The structuralist model

While the idea of postion classes can be traced back to the earliest systematic expositions of structuralist morphology (Nida 1949, Harris 1951), perhaps the clearest statement of both the method of position class analysis and the reasons for adopting such a method can be found in the more pedagogically oriented work of Gleason (1955:112):

> In some instances the number of affixes used in a single paradigm is very large; or a single word may consist of a rather long series of morphemes. It is necessary to have some simple way of stating the complex combinations which can occur. This can often be done by classifying the morphemes into groups known as *orders* which are most conveniently designated by numbers. Thus, order 1 consists of all those suffixes which can occur only immediately after the root. Order 2 consists of those which can occur immediately after a morpheme of order 1, or immediately after the root if no morpheme of order 1 is present, but never farther from the root than this. Order 3 consists of those which can occur only after roots or members of orders 1 or 2 ... Only one affix of a given order can occur in a given word. If, for example, two members of Order 1 should occur, then one would have to follow the other. By our definition, a member of order 1 can only follow a root. Our classification of affixes into orders would be shown to be erroneous, and we would have to revise it to accord with the facts of the language. Orders are, therefore, mutually exclusive classes of morphemes occupying definable places in the sequence of morphemes forming a word.

The cautionary note "this can often be done" is justified by the existence of cases in which such a strict ordering as required by the method summarized above can not be established. It is conceivable, at least theoretically, that a given root R can be followed by suffixes A and B in this order (perhaps followed by further suffixes *C, D, ...*) and the same root can also be followed by B and A (and perhaps further suffixes). In such cases, the symmetry between A and B would dictate a solution in which both suffixes appear in the same order (either order 1 or order 2), yet the forms RAB and RBA then violate the constraint of only one morpheme per order in any word-form. While generally characteristic of derivational, rather than inflectional, morphology, such cases are not at all uncommon (see Baker 1986) and we can readily cite some from Hungarian. The 'diminished volition' suffix *gat/get* and the 'causative' suffix *tat/tet* are interchangeable after verbal stems: *dolgozgattat* 'make someone work not earnestly' vs. *dolgoztatgat* 'not earnestly make someone work'. Both of these suffixes can be followed, but not preceded, by the 'permissive' suffix *hat/het* as in *dolgozgathat* 'is permitted to work not earnestly' and *dolgoztathat* 'is permitted to make someone work'.

1.1.2 The level-ordered model

In generative morphology, the primary method of describing the ordering and cooccurrence relations obtaining among suffixes is *level ordering,* one of the key ideas of Lexical Phonology and Morphology. To quote Kiparsky (1982:11)

> The model of lexical phonology directly predicts the correlation between "boundary strength" and affix order that was observed for english by Siegel (1974), and is apparently a general property of languages. The generalization is that affixes of level n are not added to stems which already contain affixes of level *n+1.*

While the level ordering mechanism of Lexical Phonology also serves other, and in some sense more important, functions, such as the organization of the phonological rules, in this brief overview these will be ignored and we concentrate on the class of morphotactic regularities that can be described by the use (or abuse) of this formalism.

It is clear that in cases such as the three Hungarian verbal suffixes discussed above no position class analysis exists while the situation can be easily described by assigning the 'diminished volition' and 'causative' suffixes to Level 1 and the 'permissive' suffix to Level 2. (In truth this is more of an illustrative example of the power of level ordering than an actual proposal. A more detailed analysis of Hungarian lexical phonology would actually assign all three suffixes to Level 1.)

At first sight it also appears to be clear that if a position class analysis exists it can be replicated in Lexical Phonology by assigning order n suffixes to level n. There is a subtle difference between the two styles of analysis: orders are built "inside out", starting with the affixes closest to the stem, while levels are really assigned "outside in", starting with the outermost affixes. More importantly, in order to replicate the constraints enforced by the position class analysis it is also necessary to rule out recursion within a level. While the theory of Lexical Phonology offers no universal constraint to rule out such recursion, it does offer a language-specific mechanism, namely morphological subcategorization frames (see Lieber 1980), which can be used for this purpose. This means that the overall generative capacity of Lexical Phonology is greater than that of position class analysis, but the increased descriptive power comes not from the level ordering mechanism but from the subcategorization mechanism.

1.1.3 Finite state morphotactics

The context-sensitive subcategorization frames commonly used in generative morphology have the power to describe any context-free (see Peters and Ritchie 1971) language *within* a single Lexical Phonological level or, what is the same, without any appeal to level ordering at all – for the case of the English morphotactic regularities that were originally used to motivate level ordering this has been noted by Fabb (1988). While position class analysis is clearly too restrictive, full context free power is probably not restrictive enough: it seems likely that e.g. Dyck languages (languages of properly matched parentheses of arbitrary depth) have no counterparts in the morphologies of human languages.

Thus there is a definite theoretical interest in theories of morphotactics that fall between these two in generative power. The most widely used such system is Kimmo Koskenniemi's (1983) model of two-level phonology/morphology, which has an explicitly finite state morphotactic component. Again, the two-level model has a number of important features, such as a declarative rule system, parallel rather than serial rule application, or the lack of subsegmental units, that must be ignored in this brief overview. And again, our interest is focussed on analyses that employ the formal apparatus but do not necessarily follow the spirit of the model. With this caveats it is clear that not only position class analysis, but also computationally inspired morphological analyses, generally depicted with the aid of boxes containing the list of morphemes and arrows showing the possible direction of attachment (for Hungarian, see e.g. Prószéky *et al* 1982) can be recast in the two-level model. As an example, in 1.2 the basic structure of the Hungarian nominal and verbal paradigms is presented using finite state morphotactics.[2]

1.2 Hungarian morphotactics

The morphological model standardly used in Hungarian generative and/or computational linguistics is based almost entirely on the pioneering work of Antal (1959, 1961, 1966). Antal, a strong advocate of the structuralist methods of linguistic description, made a clean break with the then dominant diachronic mode of theorizing and applied the best technical tools of synchronic analysis available at the time, including position class analysis. His technical contributions included the discovery of a unified system of tense/mood marking: this discovery was made possible by the systematic exclusion of the periphrastic tense/mood categories that were previously lumped together with purely morphological marking.

Once this group of four suffixes (zero for non-past, t for past, j for imperative/subjunctive and n for conditional) is identified, it is clear that they are always preceded either by the stem (order 0)

[2]For a full-blown two-level analysis/generation system see Kornai 1989

or by some combination of the three stem-forming suffixes discussed above (order 1, under the looser definition of order that permits recursion). The tense/mood group is then order 2 (no recursion) and this is followed by the portmanteau affixes indicating person, number, and direct object distinctions (order 3, no recursion). Using the formalism of regular expressions (? stands for 0 or 1 instance, $*$ for 0 or more instances, | for disjunction) we have:

$$< stem > [< gat > | < tat >]* < hat >? < tense/mood >< person/number/def\_obj >$$

In the nominal paradigm, Antal's contributions include the positional, rather than semantic, definition of case markers, and a unified analysis of the singular and plural possessive marking. While most of his analyses were adopted with only minor modifications by the subsequent generative work, the analysis of the possessive system, which on Antal's terms involved infixation, could not be properly restated in a generative setting without the rules and conventions of nonconcatenative morphology (McCarthy 1981, Marantz 1982).[3]

As a first approximation, let us ignore the infixation process and treat the possessive suffixes as portmanteau morphs indicating the person and number of the possessor as well as the number of the possessed element. In this analysis, the twelve possessive suffixes, the plural *k* and the singular zero make up order 1 (nonrecursive) which immediately follows the nominal stem (order 0). Order 1 is then optionally followed by the 'familiar plural' *ék* (order 2, nonrecursive), which is optionally followed by the 'possessive anaphor' singular *é* or plural *éi* (order 3, recursive) and finally by a case marker (order 4, nonrecursive):

$$< stem > [< possessive > | < plural >]? < fam\_pl >?[< anp > | < anp\_pl >]* < case >$$

Since a stem cannot carry both the plural *k* and the familiar plural *ék* this analysis has to be modified to exclude this combination. Unfortunately, this can not be done by reclassifying the plural as order 2, since that would permit the possessive suffixes to combine with *k*, which is also ungrammatical. Thus the part between order 0 and order 3 has to be replaced by the more complex regular expression

$$[< plural >?| < possessive >? < fam\_pl >?]$$

which makes the plural "order 1.5". Within the possessive part of the paradigm, the same method of disjunctive paths can be applied to yield the 12 grammatical combinations (singular or plural 1st 2nd or 3rd person possessor, singular or plural possessed NP):

$$([o|ai][m|d|tok])?|([u|ai][nk])?|([a][i]?)?$$

From a linguistic perspective it is hard to justify the three-way disjunction that groups together the 1sg 2sg and 2pl possessors in the first disjunct, the 1pl and 3pl possessors in the second disjunct, and puts the 3sg possessors in a third disjunct, especially as the marker of plural possession, the morpheme *i*, appears in the left-hand side of the first two disjuncts, but on the right-hand side of the third.

2 Statistical lemmatization

The most obvious thing to count in a corpus is the number of tokens for each word. Even this simple task poses some challenges for the linguistically minded researcher, because the computationally natural definition of words as maximal non-space strings delimited by whitespaces will not, as a rule, yield a reasonable classification of types. At the very least we need to remove initial and final punctuation, capitalization, and other extralexical material like (orthographic) clitics (such as the English

[3]For the details see Kornai 1989.

possessive *'s*). However, if we wish to emulate the traditional lexicographic practice we need to process the resulting strings further: we need to identify, and perhaps remove, all inflextional affixes so that word-forms can be grouped into lemmas based on stems.

For a computational linguist whose main interest is English it is not at all obvious that lemmatization of this sort is justified by anything other than adherence to lexicographic tradition. But it *is* obvious that lemmatization is a nontrivial task: simple affix stripping algorithms run into problems of overanalysis (*coed* as the past tense of **co*) and underanalysis (*hanged* but not *hung*) relatively early in their development. Since the task is hard (because it seems necessary to build a great deal of human expertise into the algorithm) and the benefits are dubious, why bother? The aim of this Section is to show, by means of statistical argumentation, that if the goal is to extend the frequency-based analysis to syntax (tagging and parsing) and/or semantics (machine translation) then lemmatization is unavoidable, at least for languages with complex inflectional morphology.

The argument is presented in two parts: first the reduction of syntax/semantics to morphology is discussed in 2.1, where it is shown that for the leading statistical paradigm, based on string matching (see e.g. Sankoff and Kruskal 1983) the complexity of the parsing, tagging, and machine translation tasks is bounded from below by the complexity of the concomitant lemmatization task. In other words, no algorithm that fails on the morphological task can possibly be successful on the syntactic or semantic tasks – a result that comes as no surprise to those sharing the traditional linguistic perception that morphology is "easier" than syntax, which in turn is "easier" than semantics.

Section 2.2 presents a general method for estimating the complexity of lemmatization, as measured in terms of training set size. To this end, the well-known statistical apparatus of absolute, relative, and adjusted frequencies (summarized in Francis and Kucera 1986:461-464) is extended by the introduction of the notions *well-represented types* and *saturated paradigms*. Section 2.3 introduces some Hungarian data that shed light on the expected complexity of the morphological task – in the light of this data it is hard to escape the conclusion that the "brute force" string matching approach is doomed to failure.

2.1 Reduction to morphology

The central idea of statistical machine translation (SMT) is to view translation as a string matching task. While it would not be fair to say that English is just badly spelled French, the remarkable success of English/French SMT (see Brown *et al* 1990) owes a great deal to the structural similarities of these two languages. Faced with pairs of sentences such as

Pétert nem kell bíztatnod
You need not encourage Peter,

the assumption that equivalent elements of the two sentences occupy roughly the same position must be strongly revised. Not only do *Peter* and *Péter* appear a full sentence-length apart, but the equivalent of the initial subject *you*, namely the suffix *-d* also appears finally. Since *need* is *kell* and *not* is *nem*, one might be tempted to conclude that the order of elements in Hungarian is exactly the opposite of their order in English. Nothing could be farther from the truth. In fact, all six permutations of the phrases *Pétert, bíztatnod*, and *nem kell* are equally grammatical, though some orders (e.g. the one given above) are considerably more likely than others.

Since a simple transitive sentence has at least 6, and a simple ditransitive at least 24 grammatically valid permutations[4] which will all be translated with the same English sentence, a conservative estimate would be that we need at least 10 times as many English/Hungarian pairs for a representative sample as we would for English/French. But this is actually not enough: if we have one order of magnitude more *prima facie* candidates for a single word, the most conservative estimate is that we will need at least two orders of magnitude more data to sort out which is the right one. Given that the largest

[4]These counts are based on the number of phrases. Unlike in true "free word order" languages like Latin, in Hungarian the order of words within phrases is quite fixed, so it would be more proper to call it a "free phrase order" language.

extant bilingual corpora are barely sufficient to derive reasonable translation statistics, waiting for a hundredfold increase in corpus sizes is simply unrealistic.

To get around the combinatorial explosion of patterns to be matched, abstract strings such as *see-PAST-2ND-PL-DEFINITE* need to be successfully matched to fully-formed words such as *láttátok*, for the simple reason that the suffix-combination *tátok* on the stem *lát* 'see' appears exactly when the verb is in past indicative 2nd person plural definite context in the English sentence. Garnering all this information from the English sentence can be tricky: for example a subject *you* informs us only that the value of person is 2nd but provides no information as to number. Nevertheless we will assume that this can be done i.e. that there is a *transfer* stage (see Brown *et al* 1991) at which English stems are tagged by morphosyntactic features that disambiguate which paradigmatic form in the lexeme of the stem is to be matched. All that needs to be said here is that the solution of the machine translation problem requires the solution of the tagging problem, (at the English side of the transfer) so the complexity of MT is bound from below by the complexity of the tagging problem.

At the Hungarian side of the transfer the cost of the matching operation can be assumed to be fixed, since the number of morphosyntactic features is strictly limited (≤ 7) and the order of Hungarian suffixes (at least of the inflectional suffixes relevant for preserving the syntactic information necessary for successful translation) is fixed. Thus the complexity of the tagging is bound from below by the complexity of the operation that assigns morphosyntactic tags to inflected word-forms i.e. by the complexity of the lemmatization (morphological analysis) task. For this task we can safely assume that matching will be performed successfully if the system is properly trained, so the real bottleneck is not the speed of the computation, but the amount of data required to perform the training.

In sum, the typological gap between English and Hungarian word order precludes translation by direct string matching and makes it necessary to apply a transfer stage that will relate inflected word-forms to morphosyntactically tagged stems. In order to do the translation we need to do the tagging, and in order to do the tagging, we need to do lemmatization. Thus the complexity of the machine translation task can be estimated from below by the complexity of the tagging task, which in turn can be estimated by the complexity of the lemmatization task. The simplest possible lemmatization algorithm is brute force table lookup: in the next section we investigate the issue of how much data we need to construct the tables.

2.2 Corpus size requirements

One method of morphological "analysis", commonly applied in English spell- and grammar-checkers, is to list every word-form in the dictionary: in addition to listing *appoint* as a verb stem, we also list *appointing, appointed* and *appoints* with the appropriate tags. Modern techniques of dictionary compression make such a direct listing quite feasible, and the savings in time (table lookup vs. runtime morphological analysis) are considerable. If a table is at hand, it can be compressed at compile time so its initial size (and the time it takes to compress it) are of little practical importance. Thus when we inquire about the complexity of morphological analysis the main question to be addressed is not how much *time* it takes to analyze a single word, but rather how much *data* it takes to create a full listing, together with the appropriate tags, for each reasonably frequent word-form. To answer this question, in 2.2.1 we introduce the notion of *well-represented* types (lexemes or word-forms) and show how the corpus size required for a given number of tokens of a well-represented type can be extrapolated on the basis of relative frequencies in smaller corpora. Next in 2.2.2 we introduce the notion of a *saturated paradigm* and illustrate the extrapolation method for English – the corresponding results for Hungarian will be discussed in 2.3.

2.2.1 Well-represented types

Let us call a lexeme or word *well represented* in a corpus if increasing the corpus size n times will increase the absolute frequency of the form by a factor of n. Suppose for some reason we need to

collect 500,000 tokens of *the*. Knowing that the relative frequency of *the* in English is almost 7% we can estimate the corpus size required for this to be slightly above 7.2m words: it might be that collecting 6.8m will be enough but we would be genuinely surprised if collecting 5m was already enough since that would mean the original estimate of 7% was seriously flawed. In general, if the absolute frequency of a well-represented element in a corpus of size N is M, and we need c copies of it, this will require, on the average, a corpus with cN/M tokens. In other words, the relative frequency of a well-represented form in a sample can be used to estimate the relative frequency of the form in the population.

How do we know if a form is well represented? Since being well represented is a property of the distribution of an element in larger samples, we cannot directly ascertain it beyond what we see in our present sample. We suspect that certain elements (those that show great fluctuations in frequency) are not well represented, and we know it for a fact that certain other elements (those grammatically correct forms that do not appear in the corpus at all) are badly represented. However, as we increase the corpus size, the fluctuations in frequency decline, and more and more elements become well represented,[5] and we can be reasonably certain that a form is well represented if, upon dividing the existing corpus into smaller segments, the relative frequencies of the form within these segments are close to one another. This closeness is generally measured by *dispersion* (see Juilland and Chang-Rodriguez 1964). If the dispersion of a form is high, this means the form was already well represented in the subsamples forming the basis of our dispersion calculations. If the dispersion is low, we should base our reverse calculations on *adjusted*, rather than on absolute frequencies.

2.2.2 Saturated paradigms

Before we can present our estimate for the corpus size required for the morphological alignment task outlined above we need one more technical term. We will call a lemma *saturated* in a corpus if each grammatically expected paradigmatic form is actually instantiated by at least a single token. For example in the Brown corpus the lemma *emerge* is saturated since all the paradigmatic verb-forms *emerge, emerges, emerged, emerging* are found, while the lemma *emanate* is not saturated since the forms *emanate* and *emanates* are both missing. In order to guarantee that a table lookup tagger will successfully find the appropriate morphosyntactic tags for any string composed of a stem and some suffixal morphemes the lemma of the stem in question must be saturated in the the training corpus.

Let us now illustrate on an English example how our reverse frequency estimate works. Consider the verb *ascertain*, which appears 11 times in the Brown corpus: 7 times in base form and 4 times with the suffix *-ed*. Since English morphology is quite simple we can be reasonably certain that the 3rd singular form is *ascertains*: chances of suppletion and/or changes to the stem are virtually nil. But let us pretend this is Hungarian where such phenomena are quite common so that we need to find out, by way of an example, exactly how the 3rd singular form looks like. Since we know that on the average 6.44% of verb-forms is 3rd singular in English, it is reasonable to assume that once we have 16 or more tokens in the lemma of *ascertain* at least one of these will be the 3rd singular form we are looking for. How big a corpus we need for that?

If *ascertain* was well represented in the Brown corpus we would only need 1.5m words, since its absolute frequency is 11/m. However, the dispersion of *ascertain* is quite low, so the computation should be based on the adjusted frequency, 5/m. Thus we arrive at the conclusion that in order to be reasonably certain that our corpus contains the 3rd singular form of *ascertain* we need to collect 3.1m words. At that point, more frequent forms such as the gerundive will, in all likeness, be also present, since on the average we get 2.44 gerundives for each 3rd singular, so at 3.1m words the paradigm of *ascertain,* and indeed the paradigm of every verb with adjusted frequency above five, can be expected to be saturated.

[5] At least in absolute terms. It is quite conceivable that the *proportion* of well-represented forms actually declines as the corpus size increases.

The main assumption here is that the choice of the stem is independent of the choice of suffix-combination or *paradigmatic slot*. This of course need not be true on a stem by stem basis: for example the frequency distribution of the suffixes will be highly distorted for stems with defective paradigms. Nor can we assume that the probability of various paradigmatic forms is constant across genres: for example the ratio of (Hungarian) imperative vs. past tense verb forms is considerably higher in literary than in scientific texts. However, as long as we control for stylistic variation (which of course impacts the frequency ranking of stems) we can proceed as follows.

In order to guarantee satisfactory performance for the K most common stems in a lexical category (for the sake of concreteness, the first 30 English common nouns) we need to determine the (adjusted) frequency f_K of the least frequent one. In the Brown corpus, this is the noun *war* with adjusted frequency $f_{30} = 0.044\%$. Since the nominal paradigm in English has only two slots, singular vs. plural, and singular nouns outnumber plurals roughly 3 to 1, $c = 4$ (out of four tokens for the stem *war* one is expected to be in the plural form) so the corpus size can be estimated at $N = c/f = 8,850$

It should be emphasized that this is only a maximum likelihood estimate: to guarantee with a high level of confidence that *wars* appears in the sample would require a considerably larger N. However, as we are interested only in a lower bound for required corpus size, the estimate $N(K) = c/f_K$ is sufficient. The question how c can be estimated for more complex paradigms will be discussed in 2.3 – here we concentrate on the relationship between N and K.

Unfortunately, this relationship between the number of stems we wish to cover and the size of the training corpus is nonlinear. To cover n times as many stems as we covered before we need to consider f_{nK} instead of f_K. By Zipf's (1949) law $f_K = C/K^B$ for some constants B, C which means $N(nK)/N(K) = f_K/f_{nK} = n^B$. As Mandelbrot (1964) observes, the linear correlation between log rank and log frequency prescribed by Zipf's law requires $B > 1$, since otherwise the sum of the frequencies would diverge. This means, in effect, that in order to double the number of stems covered we need to increase the corpus size by a factor larger than two.[6] Notice that this conclusion is in fact independent of the validity of Zipf's law. Whatever the correct statement of rank-frequency relationship might be, as long as $1/f_K$ grows only linearly with K the sum of the frequencies fails to converge.

2.3 The case of Hungarian

The database used in this study is an amalgam of three different computational efforts: the Reversed-Alpabetized Dictionary of Papp (1969), The Frequency Dictionary of Füredi and Kelemen (1989), and the two-level morphological analyzer of Kornai (1989). The Frequency Dictionary is based on 258 samples drawn from literary texts published between 1965 and 1977, totalling 508,008 tokens. Of these, only 487,450 appear in the database, because at an early stage of data entry proper names were left out.

There are 91,833 distinct words, collected in 41,884 lemmas. What is most remarkable about these numbers is that the average number of paradigmatic forms per lemma, 2.19, is so low. Is this because the corpus is dominated by words such as determiners and other function words which have only a single paradigmatic form? Among the first 50 lemmas we find 21 indeclinables which account for 119,840 forms and 29 lemmas which together contain 1,043 paradigmatic forms that account for 48,999 tokens. If we remove these 50 lemmas, which account for over a third of the whole corpus, we are left with essentially the same ratio (2.17 forms per lemma). Thus we can safely conclude that the form/lemma ratio is low not because of some high-frequency indeclinable entries but rather because the low-frequency entries show up in so few forms.

While the expected number of word-forms per lemma is much higher in Hungarian, the actually attested forms are distributed in a manner not so dissimilar to what we find in English. Thus a hardline empiricist might come to question the theory behind the original expectations: maybe the complex paradigms summarized in 1.2 exist only in the mind of the grammarians, and the actual situation, as

[6] Given the asymptotic nature of the argument it is not surprising that the effect becomes perceptible only for larger K. But for $K > 400$ both in the Brown and in the Hungarian corpus we find $f_K/f_{2K} > 2$

seen in the corpus, is far simpler. However, when we take a closer look at the data, it becomes quite clear that the supposed paradigmatic forms are all there *in the population*, even though they do not necessarily appear *in the sample*.

First of all, a handful of verbs such as the copula, *mond* 'say', and *tud* 'know' have near-saturated lemmas. Not surprisingly, these are the lemmas with the greatest frequencies, both absolute and adjusted. Second of all, the pattern of the less saturated lemmas is entirely consistent with the assumption discussed in 2.2. that a fixed percentage of verb-forms falls into the most frequent paradigmatic slot (which is *PAST-3RD-SG-DEFINITE* for Hungarian verbs), some lesser percentage to the second most frequent slot *(PAST-3RD-SG-INDEFINITE)* and so on. Even the most unlikely suffix-combination *COND-2ND-PL* receives a small, but quite definite percentage of the total, and the reciprocal of this number provides a maximum likelihood estimate of the constant c used above.

Interestingly, a reasonable estimate of c can be derived on the basis of limited independence assumptions. In the verbal paradigm, if we assume tense/mood to be independent from person/number we can calculate the probability of *COND-2ND-PL* to be $0.042 \cdot 0.0047$, which would yield a little over 17.6 *COND-2ND-PL* forms out of a total of 89,178 verb-forms: in the corpus we actually find 14. A full statistical study of the corpus is still in progress, but preliminary results indicate that the frequency distribution of suffixes in one order is surprisingly independent of the choice of suffixes in preceding and following orders.

For the verbs it makes little difference whether we base our calculation on the estimated or on the observed frequency of the least probable suffix-combination: for the verbal paradigm to become saturated we need $c = 5,066$ (or $c = 6,370$) tokens for a single stem – this has to be compared to $c = 16$ for English. If we take $K = 1,000$ (meaning we wish to guarantee saturation for the first thousand most frequent verbs) in the corpus these have adjusted frequency 8 or higher. Therefore we need to collect a corpus 630 or 800 times bigger than the present one (depending on which of the above figures we use) so we need to collect 320m to 400m words. Again, this has to be compared to the corresponding figure for English, which (based on the Brown corpus) is only slightly above 2m.

Let us now look at the similar estimate for nouns. Recall that for $K = 30$ we found $N = 8,850$ in the Brown corpus. In the Hungarian case we do not find a single lemma approaching saturation: the most frequent noun, *ember* 'man' occurs in 51 of the 714 standard paradigmatic forms (see Antal 1961), and there are only five other lemmas that contain 50 or more forms. Here again our first impulse is to revise the grammatical classification so as to bring the predicted and the observed variety of forms more in harmony. While there is no theoretical reason for doing so, in practice the sparseness of training data makes it necessary to omit the position class of anaphoric possessive suffixes: of the 111,401 noun forms in the corpus the anaphoric plural appears only once, and the singular less than 200 times. If we leave these out, the expected variety of paradigmatic forms is reduced to $14 \cdot 17 = 238$ forms.

Of the fourteen alternatives in order one, it is again the second person plural (of the plural possessive) which is the least frequent: it appears only 3 times. However, to leave it out would be completely arbitrary, as the second person singular appears over 70 times, and other person/number combinations appear even more frequently. Given our hypothesis about the orthogonality of frequency distributions along different paradigmatic dimensions, the scarcity of *aitok* and *eitek* follows from the fact that these are morphologically complex. Indeed, if we multiply out the probabilities of the various morphemes required for this suffix-combination, we get an expected frequency that is actually slightly *lower* than the observed frequency.

Since the case endings are morphologically simplex, the situation is markedly better there: even the least likely one (the 'formalis' *ként*) occurs over 180 times. The 'factive' *vá*, the 'terminative' *ig* and the 'causalis' *ért* are slightly more frequent (200 to 300 occurrences), but still nowhere near the dative, accusative, or nominative. However, as c is determined by the least frequent paradigmatic slot, these higher numbers do not figure in our calculations: combining the least frequent possessive with the least frequent case ending yields $c > 22,000,000$. For $K = 30$ this gives us the staggering result that we need over 33 billion ($3.3 \cdot 10^{10}$) tokens to get the first 30 nouns saturated.

It is worth emphasizing that this estimate is a conservative one: the possessive anaphoric forms that would add another factor of 10^4 were left out of consideration. Similarly, the frequency (both actual and adjusted) of the 30th Hungarian noun, *föld*, is 50% higher than the combined frequencies of its main English senses 'earth', 'ground', and 'soil', and indeed 50% higher that of the 30th English noun. The basic result that $N(K)$ is over six orders of magnitude bigger for Hungarian than for English holds through the complete range of K that yields reliable results at this corpus size ($1 \leq K \leq 1,500$), because the main effect distinguishing between Hungarian and English is not f_K but the difference in c ($2.2 \cdot 10^7$ vs. 16). This effect stems from the differing complexities of the nominal paradigms i.e. from a fundamental typological difference between the two languages.

3 Hidden Markov morphology

Given the size of the ongoing data-collection efforts and the amount of machine-readable material (including typesetting tapes), no more than 10^7 to 10^8 words of Hungarian will be available for statistical analysis in the foreseeable future. In the light of the results presented in 2.3 above this makes it necessary to develop a method of morphological analysis capable of analyzing fully-formed words absent from its training set. Such a method would also be desirable from the standpoint of psycholinguistics, since it is well known that humans are capable of performing morphological analysis of fully-formed words with nonce stems (Berko 1958, Anshen and Aronoff 1981, 1988).

3.1 provides a brief introduction to Markov modeling from the perspective of morphology, and sketches a design that can, at least in principle, analyze forms not in its training set. 3.2 investigates the limitations of the proposed model and tries to specify the range of expertise needs to be built into its design. In the concluding 3.3 the linguistic relevance of Hidden Markov morphology is briefly discussed.

3.1 Underlying representations and hidden states

A Hidden Markov Model (HMM) is standardly defined (see e.g. Levinson *et al* 1983) as a triple (π, A, B) where π is the *initial state distribution*, A is the matrix of *transition probabilities*, and B is the matrix of *signal distributions*. A single *run* of the model will start in some state i with probability π_i, where a signal v_j is emitted with probability $B_{i,j}$, and the model moves into state k with probability $A_{i,k}$ where another signal is emitted and so on.

The first important application of this model was in isolated word recognition, where each candidate word will have its own triple, and the recognition of a signal sequence $v_1 v_k ... v_k$ is based on computing which triple could emit this sequence most probably. In this and in later applications the full HMM model is composed of several small HMMs, each corresponding to a structural unit (word, syllable, phoneme, etc.) and these smaller units are linked together in a single network constituting a large HMM that I will call the *architecture*.

The possibility of linking smaller HMMs together to form larger HMMs i.e. the fact that the class of HMMs is *closed under substitution* provides the key to understanding their phenomenal success in linguistic applications: the smaller HMMs encode the units and the architecture encodes the tactics. To continue with our first example, the relative frequencies of word-pairs can be encoded in the transition matrix of the architecture, while the acoustic characterization of the words themselves is carried by the smaller HMMs that form, in effect, the nodes of this larger network.

While from a linguistic perspective the practice of encoding both the elements and their tactics with the same data structure might appear dubious, it is in fact this *homogeneity of encoding* that makes it possible to use global, rather than local, optimization techniques in training the network. Since global optimization eliminates hierarchical decision-making which necessarily percolates the errors of lower decisions to higher levels, HMMs have gradually displaced the more structured models in which any bad local decision leads to global failure.

Linguistic models of morphology, be they structuralist or generative, employ a strict hierarchical scheme: words are composed of morphemes, morphemes are composed of phonemes, and phonemes are composed of features. At each level, the atomic elements might be held together by complex "nonlinear" structures, and the composition of such elements will often trigger a variety of rules that manipulate these structures in various ways. Since the HMM paradigm does not permit any manipulation of the output, there can be no direct equivalent of assimilation, dissimilation, resyllabification, and other rules that pay a pivotal role in linguistic descriptions of phonology/morphology. Further, the HMM paradigm has no resources for the manipulation of different kinds of structure such as metrical and autosegmental organization or morphological bracketing because everything must be encoded in the same manner.

To see how the HMM paradigm will be applicable to morphology we will have to go back to the basic idea that there is a *hidden* process that can only be observed through the signals it emits at various stages. For our purposes, the emitted signal $v_1 v_k ... v_k$ (which is the only observable), is the appropriate *surface* phonological (or orthographic) representation of the word we wish to analyze. Leaving nonconcatenative morphology aside for the moment, the hidden process that generates this signal is simply the concatenation of the underlying affixal and stem morphemes in the appropriate order.

If the (morpho)phonology of the language were ideally transparent, this would be all that is required for the model: we build as many small models as there are morphemes, assign an output string to each, and link these smaller models together as permitted by the morphotactics of the language. Since the phonology is transparent, there is no difficulty in assigning the surface representation of the morpheme as the output string of the HMM that models it. Furthermore, if morphotactics is indeed finite state, as discussed in 1.1.3 above, the cooccurrence patterns of the morphemes can be reflected in the architecture without any trouble.

We do not need to rely on the probabilistic aspect of the output associated to the individual HMMs to reflect in the model the frequencies of the various suffix-combinations: these can be encoded in the transition probabilities of the architecture. In fact, the first order markovian assumption that the probability of a state depends on the previous state already enables us to model more complex frequency distributions than the ones employed so far: neither the independence of the suffix-distribution from the stem assumed in 2.2.2, nor the orthogonal decomposition of the distribution characterizing the suffix-combinations into one distribution per position class assumed in 2.3 are necessary for an HMM.

Finally it is clear that the model as defined so far is not restricted to word-forms appearing in its training set. In fact it is capable of correctly tagging all kinds of word-forms permitted by the morphotactics as encoded in the transitions of the larger network: as long as there is a sequence of hidden states that yields the required surface form the standard Viterbi algorithm used in HMM pattern matching will find it. Further, by using global optimization to train the architecture the reliance on human expertise can be minimized: only the output sequence (surface representation) associated with each morpheme must be handcrafted. This much is inevitable: after all, morphemes are linguistic signs so their models must, in addition to their meaning, carry some information about their form as well.

3.2 Phonological difficulties

So far we have presented a truly optimal scenario: given a list of stems in the form of surface phonological representations, and a set of affixal morphemes (lists of morphosyntactic tags associated with surface strings), we can devise an HMM that will acquire knowledge about the frequencies of various word-forms on the basis of an (unsegmented, untagged) training corpus and further, will be able to provide tagging for words not present in the original training corpus (as long as these are morphotactically correct) by standard algorithms that recover the most probable hidden state sequence. Since the role of the linguist in this process was limited to providing a lexicon (stems plus suffixes, the latter with tags) clearly the whole problem is now solved.

Unfortunately, the above scenario is based on the assumption that the phonologies of languages are transparent, while in actual fact very few languages come close to this ideal and none attain it perfectly. Let us therefore survey the range of difficulties and see how the model can be extended to cope with these. The most pervasive source of opacity is *assimilation*: for example in Sanskrit a stem ending in *a* and a suffix beginning with *i* will combine to yield *e*. Since the same vowel results if the stem ended in *e* and the suffix began with *a,* the relationship between the underlying vowel and the surface representation is to some degree opaque.

In order to solve this problem we need to revise both the architecture encoding the tactics and the smaller networks encoding the individual morphemes. The linear subnetworks corresponding the morpheme were so far assumed to have one state for each phoneme, with transitions from the last phoneme of a morpheme to the first phoneme of possible next morphemes. Whenever such a transition was running from a final *a* to an initial *e,* this now needs to be removed and replaced by a transition running from the penultimate state of the first morpheme to a single *e* state that has a transition to the second state of the second morpheme, and similarly for all other sandhi pairs.

While such a revision is theoretically feasible, in practice it would introduce so many new states (one for each relevant pair of morphemes) that it makes more sense to adopt the method standardly used in speech recognition and use *triphones* instead. (Triphones are *not* sequences of three phones XYZ – rather, they are single phones in two-sided contexts: Y /X_Z. But the term triphone is now so much part of the established terminology of the field that there is no hope for replacing it by a better one.) Triphones will still complicate each morpheme model, but will introduce only as many new states into each small HMM as there are phones, rather than introducing as many states as there are morphemes.

By using triphones *local* assimilation and dissimilation phenomena become manageable, even if the phonological rules describing these are feeding or bleeding one another. While this still leaves a number of important phonological phenomena outside the scope of the model (we will return to these in 3.3), rules of local assimilation are so pervasive that it is worth considering the class of languages where the only source of opacity is the presence of such rules.[7] Given such a language, what are our chances in recovering the underlying forms without human intervention?

Let us take a simple but typical example of the use of underlying forms in phonology. Russian has a rule of *final devoicing*: a consonant appearing in final position must be voiceless, even if it shows up voiced e.g. when followed by a vowel-initial suffix. Since there are stems in which the final consonant always shows up voiceless, generative phonology standardly captures the distinction in terms of underlying voicing contrast that gets neutralized in final position. In terms of triphone-based HMM analysis we can say that the network corresponding to stems such as *porok* 'vice, defect' ends in a single triphone K/x_y where x and y are arbitrary phones, while the network corresponding to stems such as *porog* 'threshold' splits at the penult state into two triphones $K/x\_\#$ and G/x_V where # stands for word boundary and V for any vowel.

Not only can final devoicing be expressed in the model, it can also be learned by the model in the following manner. At the outset let us equip both kinds of stems with both kinds of final triphone, and then use a Russian corpus to train the model. As a result, in stems where the final vowel is always unvoiced the transition probability to the G triphone will be trained to 0, while in the other stems it will be trained to some positive value. At least in principle we can start out with morpheme models that are full networks of *all* kinds of triphones, initially seeded by the surface value but permitted to assume other values as well.

At this point I would venture the prediction that such loose networks have so many parameters that the corpus size required for their training would be astronomical. However, it should be quite possible to tighten the model by introducing linguistic expertise in the form of pre-set parameters to the point where training on realistic corpora would become feasible. First, the probability of transitions corresponding to combinations ruled out by the morphotactics should be set to zero. For example in

[7] Dissimilation is in fact much rarer, but it comes for free once assimilation is covered.

Hungarian the probability of transition from tense/mood markers to the causative should be set to zero, while transition probabilities from the causative to tense/mood should be set to nonzero. Eliminating transitions with pre-set zeroes reduces the interconnectedness of the architecture considerably. Second, the morpheme models themselves should only contain those triphones which fit into known alternations. In the Russian example, the final consonantal triphones should only contain voiced/voiceless branches, and medial consonants need not have alternative realizations at all. This will again reduce the number of parameters to be trained considerably, perhaps to the point where training such models becomes feasible.

3.3 The linguistic relevance of the model

There is a wide range of phenomena that were ignored in the previous discussion. These can be classified under three broad headings: mid- and long-range assimilation phenomena, non-concatenative morphology, and non-monotonic rule-interaction. In Hungarian, vowel harmony is a good example of the first, and the plural possessive infix *i* is a good example of the second. But the status of the two is very different: vowel harmony is pervasive while infixation is restricted to this single example. It would require systematic changes in the basic design presented in 3.2 above to accommodate vowel harmony, but it would require only a local flattening out of the morphotactics in the manner suggested in 1.2 to accommodate the *i* infix.

The undeniable fact that certain phenomena fall outside the scope of the simple HMM framework presented above and others require counterintuitive solutions should not lead to a wholesale rejection of the paradigm. Any practicing linguist can name important theoretical frameworks that were/are the focus of much research activity in spite of well-known holes in empirical coverage. Accordingly, our conclusion is focussed not on what is wrong about the HMM paradigm but what is right about it: perhaps the most important insight a theoretical linguist can draw from the HMM approach concerns the treatment of variation.

Starting with Pāṇini, linguists take the rules of the grammar as the primary locus of variation. As Kiparsky (1979) shows, Pāṇini distinguishes three kinds of optional rules: truly optional ones, preferred, and dispreferred ones. The frequency (or at least the stylistic value) of a form therefore will be a function of the number of times optional rules of various sorts need to be applied in the course of its derivation. A related conception of Labov (1965, 1972), which forms the basis of the VARBRUL analysis now standard in sociolinguistics, is that each rule applies with a certain numerically quantifiable probability.

In the HMM paradigm there are no rules as such, and the alternations described by rules in phonology are stored in a "parallel distributed" fashion. This paradigm therefore takes the representations, rather than the rules, to be the source of variability. HMMs offer a dual method for including frequency information in the lexicon: the absolute frequency of a morpheme is determined by the architecture, and the relative frequencies of its alternate realizations are determined by the morpheme model itself.

To the extent that hierarchical structure is more important than linear structure the markovian approach to morphotactics will have to be revised. Fortunately, hierarchical structure seems to play a significant role only in derivational morphology, while the primary domain of application for HMMs is inflectional morphology (with possible extension to productive derivational affixes). Since in the inflectional domain linear order is generally fixed and arbitrary, there is every reason to believe the markovian approach will successfully account for morphotactic variability.

Whether triphone-based lexical representations can successfully account for the variability across alternative surface realizations of the same underlying form depends to a large extent on the preservation of underlying linear structure. To the extent that underlying order can be modified or changed by epenthesis, elision, metathesis, or other effects of resyllabification, the markovian approach will have to be revised. But its main characteristics, the conceptual integration of discrete and continuous variation into a unified probabilistic framework will no doubt be adopted theories of language.

4 References

Anshen, Frank and Mark Aronoff 1981. Morphological productivity and phonological transparency. *Canadian Journal of Linguistics* **26** 63–72.

Anshen, Frank and Mark Aronoff 1988. Producing morphologically complex words. *Linguistics* **26** 641–655.

Antal, László 1959. Gondolatok a magyar főnév birtokos ragozásáról. *Magyar Nyelv* **55** 351–357.

Antal, László 1961. A magyar esetrendszer. *Nyelvtudományi Értekezések* **29**.

Antal, László 1966. On the morphology of the Hungarian verb. *Linguistics* **25** 5–17.

Baker, Mark 1985. The mirror principle and morphosyntactic explanation. *Linguistic Inquiry* **16** 373–415.

Berko, Jean 1958. The child's learning of English morphology. *Word* **14** 150–177.

Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roossin 1990. A statistical approach to machine translation. *Computational Linguistics* **16** 2, 79–85.

Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra and Robert L. Mercer 1991. A statistical approach to sense disambiguation in machine translation. In *Proc 4th DARPA Speech and Natural Language Workshop*. Asilomar.

Chomsky, Noam 1957. *Syntactic Structures*. Mouton, The Hague.

Chomsky, Noam 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge MA.

Chomsky, Noam and Morris Halle 1968. *The Sound Pattern of English*. Harper & Row, New York.

Church, Kenneth W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied NLP*.

Fabb, Nigel 1988. English suffixation is constrained only by selectional restrictions.

Francis, W. Nelson and Henry Kučera 1982. *Frequency analysis of English usage*. Houghton Mifflin, Boston.

Füredi, Mihály and József Kelemen 1989. *A mai magyar nyelv szépprózai gyakorisági szótára*. Akadémiai, Budapest.

Gleason, H.A. 1955. *An introduction to descriptive linguistics*. Holt, New York.

Goldsmith, John 1992. Harmonic Phonology. In *The Last Phonological Rule: Reflections on Constraints and Derivations in Phonology*, John Goldsmith, (ed.) University of Chicago Press.

Harris, Zellig 1951. *Methods in structural linguistics*. University of Chicago Press, Chicago.

Juilland, Alphonse and Eugenio Chang-Rodriguez 1964. *Frequency Dictionary of Spanish Words*. Mouton, The Hague.

Kiparsky, Paul 1979. *Pāṇini as a Variationist*. MIT Press and Poona University Press, Cambridge and Poona.

Kiparsky, Paul 1982. Lexical morphology and phonology. In *Linguistics in the morning calm*, I.-S. Yang, (ed.) Hanshin, Seoul, 3–91.

Kornai, András 1986. Szótári adatbázis az akadémiai nagyszámitógépen. *Hungarian Academy of Sciences Institute of Linguistics Working Papers* **2** 30–40.

Kornai, András 1989. On Hungarian morphology, HAS Candidate Thesis.

Koskenniemi, Kimmo 1983. Two-level Morphology: a general computational model for word-form recognition and production, Department of General Linguistics, University of Helsinki, Helsinki.

Labov, William 1965. On the mechanism of linguistic change. *Georgetown University Monographs on Languages and Linguistics* **18** 91–114.

Labov, William 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Lari, K. and S.J. Young 1990. The estimation of stochastic context-free grammars using the Inside-Outside algorithm. *Computer Speech and Language* **4** 35–56.

Legendre, Géraldine, Yoshiro Miyata and Paul Smolensky 1990. Harmonic grammar – theoretical foundations, University of Colorado at Boulder Institute of Cognitive Science Technical Report # 90-5.

Levinson, Stephen E., Lawrence R. Rabiner and M. M. Sondhi 1983. An introduction to the aplication of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal* **62** 4, 1035–1074.

Lieber, Rochelle 1980. On the Organization of the Lexicon, PhD Thesis, MIT.

MacWhinney, Brian 1992. . In *Approaches to Hungarian*, István Kenesei, (ed.) vol. 4, .

Makhoul, John and Richard Schwartz 1986. Ignorance modeling. In *Invariance and Variability of Speech Processes*, Joseph S. Perkell and Dennis H. Klatt, (eds.) Lawrence Erlebaum Associates, Hillsdale, NJ, 344–345.

Mandelbrot, Benoit 1964. On the theory of word frequencies and on related markovian models of discourse.

Marantz, Alec 1982. Re Reduplication. *Linguistic Inquiry* **13** 435–482.

Markoff, A.A. 1913. Essai d'une recherche statistique sur le texte du roman 'Eugene Onegin'. *Bull. Acad. Imper. Sci. St. Petersbourg* **7**.

McCarthy, John J. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry* **12** 373–418.

Nida, Eugene A. 1949. Morphology. In *University of Michigan Press*. Ann Arbor.

Papp, Ferenc 1969. *A Magyar Nyelv Szóvégmutató Szótára*. Akadémiai, Budapest.

Prince, Alan 1992. Remarks on the Goldsmith-Larson Dynamic Model as a theory of stress. In *Rutgers Center for Cognitive Science Technical Report*. no. 1.

Prószéky, Gábor, Zoltán Kiss and Lajos Tóth 1982. Morphological and morphonological analysis of Hungarian word forms by computer. *Computational Linguistics and Computer Languages* **15** 195–228.

Sankoff, David and Joseph B. Kruskal 1983. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Addison-Wesley.

Schabes, Yves 1991. *An inside-outside algorithm for estimating the parameters of a hidden stochastic tree-adjoining grammar*. ms, University of Pennsylvania.

Siegel, Dorothy 1974. Topics in English Morphology, MIT, PhD dissertation, Cambridge MA.

Zipf, G.K. 1949. *Human behavior and the principle of least effort*. Addison-Wesley, Reading, MA.