University of Oslo February 1992

A MODEL FOR A NON-DISCRETE GRAMMAR

by

John Grøver

Acknowledgements

In the course of this work I have particularly profited from the help and support from my advisor Even Hovdhaugen. This study would not have been the same without his suggestions, interest and encouragement.

I am also grateful to Hermann Ruge Jervell for discussions on syntactic dependency and for his suggestions to the syntactic functions.

Among the large number of other people who have contributed through their help or through discussions on the present or related topics, I would like to mention Manjula Arulchelvam, Steinar Bjerve, Olav Gjelsvik, Jørund Gåsemyr, Károly Halász, Nils Lid Hjort, Stig Johansson, Knut Omang, Júlia Pajzs, Ferenc Papp, Else Ryen and Erik Torgersen.

As I collected the material for the text corpus in Budapest, I was especially grateful for the help I received from Lajos Kiss at the Lexicographic department at the Academy of Sciences as well as from Ödön Drobek and Éva Kissné Szabó at MAHIR, Gyula Nagy at the periodical Hitel, Iván Érsek and Vibjørn Madsen at the newspaper Üzlet, and András Bencsik at Pesti Hírlap. My special thanks to the typographer Tamás Nagy for removing the typographical codes from the newspaper texts.

For the project, which originally started out as an investigation of the morpheme order in Hungarian, I was granted a scholarship from The Norwegian Council for Research in the Humanities, for which I would finally like to express my gratitude.

Contents

INTRODUCTION

Grammar is traditionally discrete, in that it presupposes a segmentation of the acoustic speech sound sequence before it handles the segments as discrete units. This conception penetrates all levels of linguistic analysis: Phonology deals to a large extent with the segmentation process by reducing the large variability in the acoustic signals to a very small number of segments. In morphology, the major problems through the last decades have been centered around the segmentation of words. Syntax normally handles independent, discrete units, whether these be words or morphemes or whatever their status.

But the acoustic signals are not discrete, and it is also difficult to propose that the mental content arising from the interpretation of speech should be discrete. Much of the work within cognitive science in recent years (Lakoff (1987), Taylor (1989)) has pointed towards semantic categories as being of a basically fuzzy character, without clearcut borders between the one category and the other. The limited applicability of semantic feature theory and generative semantics point in the same direction. Thus, if concepts are basically prototypical, then even the semantics of language will be basically non-discrete. If we also assume that grammar is the systematical link between the acoustic signals of speech and their semantic interpretation, then we are left with the somewhat strange situation that both the input to and the output from the grammar are (at least partly) continuous while the grammar itself is discrete.

There may be several reasons why grammar has developed in this direction. There is a strong historical tradition, stemming back to Greek antiquity, in considering language as composed of discrete words or other units. Also, the invention of the mathematical tools and the computational power needed for handling continuous phenomena is, in this context, relatively new. There is also a very strong historical tradition within scientific methodology of establishing discrete units as the basis for scientific description. Finally, - and needless to say - everyone agrees that discrete units do, in some form or other, and to some extent or other, exist in language: We may well isolate a word and consider it independently of its context. The presence of such discrete units does, though, not necessarily imply that the grammar itself has to be discrete.

The present study will briefly look upon some of the reasons why grammar is traditionally considered discrete. The main point is, though, to suggest a model for a non-discrete grammar and present a pilot investigation of what can be found to be some of the non-discrete syntactic properties of Hungarian speech sounds.

CHAPTER 1: DISCRETENESS IN LINGUISTICS

In all grammatical models, the problem of how continuous data are transformed into discrete units is of fundamental importance. In the branch of phonology, the sound stream is segmented into discrete, successive units. Much phonological theory has dealt with the relation between phoneme inventory and sets of rules and the question of how to constrain and evaluate these, but the segmentation process itself is hardly questioned. The core function of phonology is to be found in the segmentation process and how to identify a stretch of speech sounds as a discrete unit. If the axiom of discreteness is removed from linguistic theory, then it is probable that phonology as a separate branch will disappear as well.

The same may be said about morphology, which is ultimately concerned with segmentation problems. The aim of morphological models through the last decades has been to identify discrete units and relate them by rule in word-formation. The problems for the rule definitions stem from the difficulties in segmentation: Either there is no clear phonetic boundary between two units, or there is no clear boundary between the categories exposed in a segment. In either case, morphological theory is introduced to handle the problem, and the aim is traditionally to establish a one-to-one relationship between discrete grammatical categories and discrete surface strings. A grammatical model which does not presuppose discreteness will not be in need of a particular morphology either.

Syntax will propose discreteness of the units (or try to establish such discreteness) to the extent that it succeeds in becoming an independent system. This is reflected in the view that a syntax is satisfying to the extent that it can account for the data without exceptions to the rules. This is normally considered to lend explanatory power to the theory. The ideal syntax is, at least within some schools of thought, the one which can account principally for all grammatical sentences in a language with a minimum of ad hoc-solutions. This is just another way of saying that the syntax should behave as an independent system. In this sense, syntax proposes discreteness to the extent that it tries to become context-free.

A context-free grammar over natural language is shown to be impossible. Context-dependency appears in cases when there is a dependency between two units which cannot be accounted for by a general rule, i.e., when it will be necessary to introduce lexical information in the syntactic interpretation. A maximally constrained (and, it is sometimes argued, maximally successful) syntax reduces the amount of lexical information required to a minimum. This will also be the syntax which is maximally independent of its units, i.e., which has a maximal autonomy. The extreme case of a syntax functioning entirely without lexical information, i.e., which takes random lexical units as input, would imply a maximal degree of segmentation.

It has, though, no practical interest for natural language. What is normally considered a minimally successful syntax will be the one which lists all possible word combinations, i.e., all possible sound strings, in a language (say, for practical reasons, below some upper limit for its length). This kind of syntax would be characterized by refraining from the segmentation: It recognizes no words or separable units in the sound stream. It presupposes no boundaries neither on the sound nor on the meaning level. And it makes no generalizations on the composition of the sound sequences.

Thus, the more successful - in the sense of capturing generalizations on dependencies between units - a syntax is, the more profound will the inherent segmentation be. In view of this, even syntax in its modern form can be seen as ultimately concerned with problems related to discreteness and segmentation: If the discreteness resulting from the design of the syntax is proportional with the successfulness of the syntax, then evidently the lack of discreteness in language is what constitutes the major topics for syntactic theory.

The more or less explicitly defined aim of much current syntactic theory - to achieve autonomy and generality -is thus an enterprise which is intimately connected with the discreteness (or the lack of it) of the linguistic units.

The extreme case of an entirely autonomous syntactic module leaves us with a lexicon of absolutely discrete items, which can have any form and any content, and which are segmented once and for all.

Thus when syntax strives towards independence from lexical information (or, better, tries to reduce this information to a small set of qualities which can distinguish the items, such as word class information etc.) in order to become autonomous, it moves towards an inherent definition of the lexical units as discrete.

But what about a syntax which relates units on formal criteria? This could theoretically be independent of lexical information. If we imagine a language in which all words could be uniquely determined as to word class on formal (morphological) criteria, we could set up a syntax which could function, at least to some extent, independently of the lexicon.

It is another important property of natural language that such languages do not exist. This kind of formal criteria for lexical selection characterizes artificial languages, but no natural language has been found which exposes a sufficient regularity to allow for purely formal selection criteria. And even if a language were found in which broad word classes could be determined unambiguously and exclusively on formal criteria, we would still find that there would be extensive collocational restrictions and idiomized expressions which could not be captured and determined on formal criteria. This means that even if all word class characteristics can be determined unambiguously on formal criteria, the syntax will still not be context-free, and cannot do without lexical information.

When lexical entries are represented in syntactic literature as 'sleep [V]', 'house [N]', the syntactic lexical information called word class identification can be seen as a compensation for the lack of such formal morphological marking on the phonetic surface. In a language which has formal exposition of word class, it is not necessary to add this information in the internal lexicon, since it is there on the surface. Thus the lexical selection restrictions are compensations for insufficient formal marking.

Therefore, if all distributional interdependencies between syntactic units were formally unambiguously marked, then the syntax could achieve true autonomy. But how much formal marking would be needed for this? The distributional interdependencies between syntactic units are so rich and complex and idiosyncratic, that the formal marking would necessarily have to be so extensive that it would in itself amount to a full syntactic description of the language. This description would, though, not be in the form of a constrained, generalized syntax, but, on the contrary, in the minimally constrained form of a listing of acceptable strings.

Thus syntax cannot become independent (in the sense of functioning by general context-independent rules) by means of selecting on formal criteria. The more general a syntax is, the more will it assume discreteness of the signs. And since syntax as a subdiscipline of linguistics necessarily must be concerned with generalizations, even syntax is ultimately concerned with this presupposed property of natural language.

Discreteness and arbitrarity.

We therefore conclude that linguistics, such as it is normally conceived in modern linguistic tradition, is intimately connected with the conception of language as composed of discrete units.

The discreteness discussed so far pertains mainly to the segmentation of units in linear succession, i.e., a horisontal segmentation of the continuous acoustic signals. Speech is, though, no less continuous vertically. At a given point of time, an acoustic signal can have any degree of intensity at any perceptible frequency. This gives an immense variability to the acoustic spectrum, and within the limits of articulatory constraints, speech sounds can vary almost infinitely. It is the task of phonology in discrete grammar to reduce this huge variation potential to a small number of discrete phonemic units, i.e., to carry out the vertical segmentation. A model which assumes arbitrarity must also assume that an acoustic signal corresponds to either the one meaning or the other. It is systematically alien to the Saussurian model that a slight change in form can imply a slight change in meaning. Two similar sound sequences do either signify the same thing, or they signify two things which - by the arbitrarity - in principle can be completely different. "Si le signifiant et le signifié étaient continus, il faudrait qu'à des signifiés voisins, c'est-à-dire partiellement différent, correspondent des signifiants voisins. [...] La correspondance entre les deux ne serait donc pas arbitraire." (Mandelbrot 1954).

The content of this is that the sounds over a continuous spectrum must be grouped into a limited set of discrete classes. Sounds which signify the same thing belong to the same class, those which signify different things belong to different classes. It is in this respect irrelevant for the model whether the meanings signified are 'similar' or not: Within the approach, there are only two values - identity or difference. There cannot be 'identity to a smaller or larger degree'.

The arbitrarity as a theoretical principle thus means that the segmentation into discrete units is carried out vertically as well as horisontally. The infinite number of possible phonetic realizations which constitute the paradigm over a part of a syntagm is chunked into a limited number of discrete symbols.

But the very principle which prohibits gradual signification change by gradual vertical change from one sign to another is of course no less valid for the horisontal dimension. Except for distinctive opposition between long and short phonological segments, post-Saussurian linguistics has not recognized a gradual semantic difference between, say, fast and slow speech. The actual content of this is that grammatical models do not consider time as a parameter for linguistic variation, but has reduced the horisontal dimension to a matter of order only.

Therefore, there is an intimate and systematical connection between the conception of language as composed of discrete units and the modern linguistic axiom of the arbitrarity of the linguistic sign.

Consequences of discreteness.

A consequence of the discreteness (the arbitrarity) is that grammar cannot account for gradual signification. An utterance may be said in a kind or an angry voice, or it can signal irony or sarkasm, contain a slight question or a slight doubt etc. This is gradual variation which highly influences the meaning of an utterance, but it cannot be accounted for within a grammar based on arbitrarity in the signification. Grammatical theory therefore excludes such continuous signification from the linguistic sign proper, and narrows down the meaning of the units to a timeless concept of a referent. All non-discrete processes of signification are relegated to the social space, to be accounted for by the sidebraches of sociolinguistics, psycholinguistics, historical linguistics, justified by the distinction between langue and parole.

This constraining and delimitation of the signification reached its peak with the theories of semantic features, by which not only the expression units, but their meanings as well, are recognized in a system of binary oppositions. This is a natural endpoint for a systematic approach based on arbitrarity: Not only the sounds, but even the meanings are segmented, such that all 'allo-meanings' belonging to one meaning segment point to the same concept of a referent.

The basic content of discreteness in linguistics is the reduction of structural variation over a certain interval, a levelling which permits us to handle the entire interval as a unit with no internal structure, and to identify it with a label. Clearly, if we consider the possible variation over the entire range of audible frequencies in the span of, say, a morpheme in syntax, this vast variability, with a possibly very complex structure indeed, is reduced to one single integer item, from which one single branch is drawn in the syntactic tree. The reduction (i.e., the simplification) is enormeous, and the vast amount of information contained in an utterance is reduced to a simple tree structure with a small number of branches. Evidently, the equally vast amount of semantic information which this utterance may transmit will be correspondingly reduced in a model which tries to account for the semantic interpretation in terms of the distinctive potential in this simplified syntactic structure.

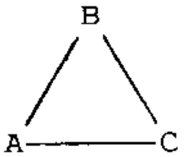
Another important consequence of grammatical discreteness is that one part of a unit cannot be related to a part of another, say, the end of one morpheme and the beginning of the following. But there are hardly any languages where morpheme boundaries are absolute, in the sense of being devoid of phonetic interdependencies between morphs. Turkish is often referred to as optimal for a morphemic description, but even here we find clear consonant assimilation processes over morpheme boundaries, in addition to the fundamental vowel harmony which penetrates the grammatical system.

In short, discreteness in linguistics has the consequence that a very large part of the structure in speech sounds remains unrelated to the semantics of speech, and subdisciplines (phonology, morphology, sociolinguistics etc.) are established to account for the variation.

The arbitrity of the Saussurian sign.

Saussure argues for the arbitrarity of the linguistic sign in Cours p.97ff. We quote Emile Benveniste: "[Saussure] entend par "signifié" le concept. Il déclare en propres termes (p.100) que "le signe linguistique unit non une chose et un nom, mais un concept et une image acoustique". Mais il assure, aussitôt après, que la nature du signe est arbitraire parce que il n'a avec le signifié "aucune attache naturelle dans la réalité". Il est clair que le raisonnement est faussé par le recours inconscient et subreptice à un troisième terme, qui n'etait pas compris dans la définition initiale. Ce troisième terme est la chose même, la réalité. [...] Il y a donc contradiction entre la manière dont Saussure définit le signe linguistique et la nature fondamentale qu'il lui attribue. [...] Entre le signifiant [= image acoustique] et le signifié [= concept] le lièn n'est pas arbitraire; au contraire, il est nécessaire". (Benveniste 1939).

Clearly, what Benveniste points to is a lack of distinction between two of the corners in the by now traditional 'triangle of signification':



Corner A denotes the sound form of, say, a lexeme. B denotes the mental content associated with it - the concept attached to it, the thought, or, as some will say, the meaning of the sound, while C is the thing in the world referred to by the lexeme, now commonly called the 'referent'. Benveniste simply says: The relation AC is arbitrary, but the relation AB is not. Has Saussure failed to note the distinction between B and C, the concept and the referent?

This can hardly be the case, when compared with the opening lines of "Principes géněraux" in CLG p.97: "Pour certaines personnes la langue, ramenée à son principe essentiel, est une nomenclature, c'est-à-dire une liste de termes correspondant à autant de choses". Saussure rejects this nomenclatural conception of language (where the units are names for things in the world) as simplified, and it is one of the major concerns in the course to show the insufficiency of the conception of language as a nomenclature. It is essential to Saussure that language does not function as a series of labels, but that both in the content and in the form will each part be inextricably connected to the whole and the whole to the part.

He frequently states that thoughts are formless if they are not supported by the structuring presence of language. "Il n'y a pas d'idées préétablies, et rien n'est distinct avant l'apparition de la langue" (CLG,p.155). "La pensée, chaotique de sa nature, est forcée de ce préciser en se décomposant. Il n'y a donc ni matérialisation des pensées, ni spiritualisation des sons, mais il s'agit de ce fait en quelque sort mystérieux, que la "pensée-son" implique des division et que la langue élabore ses unités en se constituant entre deux masses amorphes" (CLG,p.156, quoted in Benveniste). The conceptual structure is not independent of language, it is not there a priori. And the linguistic units are not given a priori either: The sounds of speech are subdivided into distinct units only by the presence of the concepts of mind. This clearly tells us that there is a high degree of interdependence between the concepts and the acoustic images, or: between the signifier and the signified. This is, as Benveniste points out, not an arbitrary relation. It is the very opposite: It is necessary, since the signified owes its presence and form to the signifier and vice versa, and if one of them had been differently segmented from the amorphous substance it arises from, the other one would have been different as well. This is fully in line with a cornerstone in post-Saussurian structuralism, that the part depends on the whole, and the whole on the part.

According to this view, therefore, the relation between the corners A and B is not arbitrary, but the relation between A and C is. For A and B, they are inextricably interdependent, and a change in one of them may give a change in the other as well. This is well known from historical linguistics, and it is known in the everyday experience that the way a sentence is uttered heavily influences its meaning. The word 'yes' can be pronounced with an infinite range of meanings, including 'no'. If the relation between the sound and its meaning were indeed arbitrary, this should (according to the above) not be possible. In a truly arbitrary relationship, the meaning cannot change gradually by a gradual change in the form.

Since modern linguistics evidently rests on the presupposition of the discreteness of the signs, or in its methodology tries to impose discrete segmentation on the speech sounds, the relationship between the corners B and C should be important for linguistics. For the very reason that the two corners can be difficult to distinguish, the problem can be difficult to delimit precisely.

How is it possible that Saussure can have these things confused, when he explicitly states that the sign proper is to be found in the relationship AB?

One historically motivated explanation could be that it is in fact not Saussure, but rather Bally and Sechayaye, the editors of CLG, who have the things confused. In their answer (together with Frei) to Benvenistes article in Acta Linguistica (1939), they are not willing to understand Benvenistes critique, and do in fact seem reluctant to accept a principled difference between the meaning and the referent (the corners B and C in the triad of signification), at least not in the same way as Benveniste does.

There is an extensive terminological confusion among semanticists in matters pertaining to the three corners of the triangle of signification (Lyons 1977). This confusion may be conceptual and not only terminological. Imagine how difficult it will be to explain to a child the difference between a house and the meaning of 'a house'. This conceptual difference is not anything which necessarily and immediately presents itself to any speaker of a language, but may be a highly culturally conditioned philosophical distinction.

There is no need to go deeper into the distinction here, but

I will only tentatively suggest the following definition of the three corners of the triangle:

A. By the sound form is meant the very physical appearance of language: the sound waves. It is a purely physical object.

B. The second corner represents the interpretation of language. It refers to purely mental phenomena of interpretation, taking place in the minds of the speakers.

C. The referent, as the object or state being pointed to, may of course be either physical or mental or any other state, but is distinguished from the second corner by being pointed to.

Much of the terminological confusion pertains to the distinction between the latter two. A sense, or even a sound form, may of course be pointed to and thus behave as a referent, and when the referent is a mental state or when some idiomized phrase functions as a 'word form', or, even worse, when one talks about the 'referents' of grammatical particles, or the 'referents' of prosodic phenomena, the problems for precise distinctions become acute. What is the referent of the word 'too'? Obviously, there will be many cases where it is very hard to make any clearcut distinctions between the sense and the referent.

If, therefore, we assume that the distinction is sometimes easy to make explicit, and sometimes it is very difficult to do so, we may perform a quick jump in our argumentation and state that the linguistic sign is sometimes arbitrary and sometimes not.

It will be the relation between the referent and the signifier which is arbitrary, and not the relation between the acoustic impression and the mental image it evokes. It is in this context of importance to note that language may be arbitrary and discrete in its naming function, when it is pointing to extralinguistic phenomena, but not in its grammatical functions. We may well establish an intuitive understanding of the referents of nouns such as 'house' and 'tree', but most people will probably have more problems in grasping the fundamental difference between the referent PLURAL and the meaning PLURAL, or this difference for, say, such grammatical categories as TENSE, ASPECT, VALENCE, CONJUNCTION, NEGATION etc. These do, though, constitute the core of grammatical systems, and it may indeed seem paradoxical that the significational nature of these linguistic units shall be conceived as arbitrary if it is the relation AC in the triad of signification which is characterized as such, and even more so if the discreteness of grammar is systematically linked to the arbitrarity of linguistic signs.

Clearly, the arbitrarity will be a valid axiom for a nomenclatural conception of language, and not for the components of a sign functioning such as Saussure describes it. Seemingly paradoxically, Saussure ends up with arguing for what he was about to reject.

There is, though, probably no other solution for a model presupposing a grammar functioning on discrete units. It is a return to language as nomenclature, in which linguistic units point by arbitrary symbols to referents in the extra-linguistic world. This situation in linguistics must be traced to the fact that the discreteness of the units, situated in the realtionship between the sound and the interpretation, is theoretically motivated by the arbitrarity valid for a completely different relationship. It is arbitrary that a horse is named 'horse' in English and 'equus' in Latin, but the relationship between a name and a referent does not enter into the syntactic interpretation of language. Grammar is situated as an interpretative module between the sound and its meaning, it is not found between the sound and the referent. The proposed arbitrariness in the naming function can therefore not motivate a grammar which presupposes a segmentation of the sound stream into discrete units.

This situation is, though, very different for a sign of the motivated sort which it is Saussure's intention to outline, and which Benveniste terms necessary. For this sign, the segmentation brings structure to both sound and meaning, and it is an integrated part of the grammar, but now the segmentation is no longer systematically necessary.

Thus: If the relation between signifier and signified is arbitrary, then a segmentation is systematically necessary, and the signs must be discrete. If, however, the relation is necessary and motivated, then the signs need not be discrete.

The theoretical implications of this is first of all that we can investigate the non-arbitrary signification in language by setting up a grammar in which the signs need not be discrete: They may be more or less formally divided, a word or a morpheme boundary can be more or less prominent along a continuous scale.

In other words, since there are no discrete segments in this grammar, there will be no lower limit for the extension of the stretches of sound which are related in syntax. The theoretical possibility of infinitely small segments amounts to infinitely large computational requirements for the syntax. In reality, therefore, we can assume that there must be a minimal limit for the 'non-discrete' symbol. There is a level for what can be perceptually discriminated, and this level could possibly constitute the size of a 'segment' in non-discrete syntax.

What corresponds to larger segments in a discrete grammar (such as morphemes and words), will in the non-discrete grammar have an internal syntactic structure. This structure is exactly what makes this stretch of sound non-arbitrary.

Therefore, a syntax over non-arbitrary linguistic signs cannot ignore their internal composition. We must assume that all physical properties of the signs can have an impact on a non-discrete syntax. This means that the phonetic properties of sound frequency distribution, intensity and duration, e.g. along scales where the perceptually discriminable is the measurement unit, will be possible candidates for parameters to a non-discrete syntax over non-arbitrarily signifying speech sounds.

The form of non-discrete grammar.

A non-discrete grammar will take the physical properties of speech as parameters. The question is, then, how these shall be related to create a syntactic interpretation of a syntagm.

The dependencies with a significational function will typically appear in the co-occurrence of events: A thing co-occurs with a sound which becomes the name of the thing. A thought accompanies a thing or a sound (the name of the thing). A thing co-occurs with another thing, and the one will 'stand for' the other in a metonymical relation. One sound co-occurs with another sound in phonological structure. We need not ask what comes first, the utterance or its sense, the utterance or the referent. From a descriptive point of view, it is the co-occurrence, i.e., the proximity in space and time, which makes signification possible.

What makes a structure linguistic is the presence (for oral language) of speech sounds as semiotic events. Signification takes place in the relation between sounds and non-sounds, and between sounds and sounds (the grammatical system). It is in principle not necessary to distinguish too sharply between these relations: The co-occurrence of a sound and a connotation, a sound and a thing (a referent), and a sound and another (neighbouring) sound are all relations in the structure which makes the linguistic system meaningful. The structures of phonology need - from a semiotic and in particlar from a non-arbitrary point of view - not differ principally from the structures of semantics, nor from the structures of non-linguistic signification. The dependencies which constitute grammatical structure will be between sounds and sounds, while the dependencies which constitute semantic structure will be between sounds and other things (thoughts, things, events). The dependencies which constitute conceptual structure will be between things and things (although these things may of course be speech sounds as well).

Thus signification, defined in this way, will emerge from a distributional structure, which is often described in probabilistic terms. From our point of view, of central importance is the fundamentally non-discrete nature of probabilistic rules. In an algorithmic rule system, a rule either applies or it does not apply: There is no third alternative, and there is no gradual transition between the two rule applications. In syntax, there is either a branch or there is not a branch. An algorithmic rule is therefore fundamentally discrete, and it requires discrete input and it produces a discrete output.

Although a probabilistic rule must also identify its input in some way or other, the dependencies which it assigns to the input is of a gradual character. It basically involves the dimension of time in a way which algorithmic rules do not. High probability means 'often-occurring' and thus short intervals along the time line, while low probability implies that there is much time between each occurrence of the input. As was argued above, the horisontal discreteness in language is connected with the exclusion of time as a relevant parameter for syntax. When grammar recognizes only the order (and not the distance in time) between vertically discrete symbols, then it becomes horisontally discrete. In this sense, an algorithmic rule system is appropriate for discrete grammar, but for a non-discrete grammar, which takes time as a parameter for syntactic dependency, a probabilistic rule system is more appropriate, since its basic measurement is the time interval between occurrences of symbols. Therefore, if syntactic dependency is defined probabilistically, this means that time is a parameter for the syntax.

If we define non-discrete dependencies by means of probabilities, it will be the over- or underrepresentation of one entity in the environment of another which establishes a significational relationship between them. We can establish such dependency between sounds and other sounds in a syntactic distributional structure, or we can see it as occurring between sounds and non-sounds in a more specificly semantic distributional structure. The latter dependencies, by which the signification as assignment of semantic meaning can be seen as taking place in the over-representation of some sounds in the presence of some things, will not be considered in the present study. For a discussion of both kinds of distributional structure, see e.g. Harris (1955).

Conclusion.

We have argued that one of the most prominent characteristics of modern linguistics is the presupposition of discreteness, which is systematically connected with the theoretical axiom on the arbitrarity of the linguistic sign. This also tells us that many of the major problems in linguistics ultimately depends on this preconception. If we remove the axiom on arbitrariness from linguistics (which there is now is even some empirical evidence for: see e.g. Bybee 1985), we will find that we cannot immediately assume that there are discrete boundaries between the various parts of syntagms. Rather, we must assume that boundaries can be more or less prominent, along a continuous scale. A grammar which will capture these features of language must be a non-discrete grammar which takes minimal units as input.

We have outlined the traditional kind of grammar as a discrete grammar. It is characterized by proposing a limited set of units upon which a limited set of rules apply. The units are discrete both horisontally and vertically, which means that neither the rules nor the set of units are affected by time considerations, and all prosodic phenomena are excluded from the discrete grammar proper. Except for such cases as distinct quantitative opposition in phone length, a discrete grammar recognizes only the relative order of constituents, and not their actual duration. This is a necessary consequence of the theoretical assumption of the arbitrarity in the signification of the units.

What characterizes this kind of grammar can also be stated negatively: It lacks any notion of time, it does not approve a probabilistic interpretation of correlation, and the physical form of its units is not relevant to it. It thus excludes all perceptual processes from the interpretation of speech, and sees the mapping from phonetics to semantics as a purely logically based interpretation.

If perception and logic are both considered parts of cognition, then we could say that discrete grammar is situated at a fairly high cognitive level, and does not include the more basic cognitive processes in the grammatical competence.

The alternative to this traditional model will be a grammar which 1. measures speech in the dimension of time, 2. assumes non-arbitrary and non-discrete signs (i.e., in actual fact, minimal segments, preferably smaller than a lower limit for perceptual discriminability), and 3. allows for probabilistic rules.

A grammar along these lines will to some extent share the characteristics of perceptual processes on sensory data, and may be given a form which makes it possible to describe (at least parts of) grammatical interpretation as a matter of perception.

Since this kind of grammar does not operate on larger segments, and since it furthermore is obvious that (to a larger or smaller extent) segmentation in language is possible, it is reasonable to assume that the segmentation may be part of the output from the non-discrete grammar. The segmentation process must belong to the interface between syntax and semantics, which is what we expect if there is a non-arbitrary relationship between the sound and the meaning: The sounds must be segmented and syntactically related in the same process which assigns meaning to them.

There will be no need for a particular phonology or a morphology as part of the non-discrete grammar proper, since the function of these components is mainly to provide a discrete grammar with convenient segments.

A non-discrete grammar will thus consist of the three parts:

1. Phonetics.
2. Syntax.
3. Semantics.

The phonetics will deal with the identification of sounds, and will be closely connected to acoustic perceptual processes.

The syntax will consider the sound forms as continuous, and will investigate dependencies between the phonetic parameters of frequency, intensity and duration.

The semantics must account for how the syntactic structures give cues to the interpretation of the phonetic signals, and how segmentation (into words, phrases etc.) can be carried out.

A lexicon may be an output from this grammar. A phonology of a language for a discrete language, i.e., the basic vertical segmentation, and a morphology over the basic horisontal segmentation, may also be part of the output from a non-discrete grammar.

There will not be any necessary incompatibility between a discrete and a non-discrete grammar. It is fully possible to consider these as two different analyses over the same data, designed to solve different problems, and being characterized by having different ranges for the phenomena they can account for. It is also possible to see them as models of co-existing grammatical competences in native speakers, where the non-discrete grammar provides the discrete grammar with a lexicon of linguistic units, such as phonemes, morphemes, words, phrases, as well as a possible basic non-transformational grammatical component. A non-discrete grammar may also account for a large number of problems concerning semantic interpretation which a discrete grammar is systematically prevented from being able to handle (signification in general, and in particular signification pertaining to prosodic or other continuous phenomena). Thus, to some extent, a non-discrete grammar can be seen as providing the input to an optional discrete grammar, in addition to functioning as an independent interpretative module. The status of both grammars and their function within the language system will be discussed in chapter 3 below.

CHAPTER 2: NON-DISCRETE PROBABILISTIC GRAMMAR

A basic-level grammar will be determined from the interdependencies in continuous acoustic data. We will here consider some possible approaches for studying such interdependencies, and report a pilot investigation based on some of these.

As was mentioned above, continuous grammar should take acoustic parameters as input, and compute the interdependencies between speech sounds from the distribution in a corpus. This corpus should, needless to say, consist in a sufficiently large sample of speech, in which the co-occurrence probabilities of the sounds are measured. In the below investigation, a text corpus has been the basis for the measure of the distributional properties, and this puts strong constraints on the investigation, since it makes a measurement of the acoustic parameters difficult or impossible.

A fundamental problem for the investigation concerns how we identify and measure the sounds. One of the main goals for a non-discrete grammar lies in the identification of speech sounds not as a succession of discrete symbols, but rather as an interconnected network of a large number of acoustic parameters which can have values on a continuous scale, and for which the distance between the measurement points in the corpus is as small as possible. This is probably as close as we can come to a truly continuous description. The present investigation does, though, not contain a sound definition of this sort, and will, for mainly practical reasons, be based on a traditional symbol identification procedure. It will, though, still imply a much better approximation to continuousness (as compared to traditional grammatical investigations) by its dependency measurements across grammatical borders and by its small distance between horisontal measurement points.

A non-discrete description should principally approximate continuousness along all the dimensions it involves. There are at least three such dimensions: 1) The dimension of time distance between sound occurrences, 2) The dimension of sound frequency, 3) The intensity of sounds. In each of these dimensions, a resolution, i.e., the distance between measurement points, must be determined.

For the horisontal dimension, the segments (i.e., the interval between two measurement points) must of course be much smaller than some average "phoneme length", which can be set roughly somewhere around 100 milliseconds. What ultimately will be determining for the horisontal resolution is the computational cost. We are concerned with investigating the distributional properties of the sound segments in a corpus, and the smaller these segments become along the time axis, the more such segments will there be. Also, when we subsequently will investigate the properties of syntagms (for example, chosen from the corpus), the computational cost of this will rise rapidly with increasing fineness in resolution. The lower limit for the horisontal resolution may thus be partly determined by the computational power available. Evidently, the computational cost will approach infinity when the resolution approaches zero.

On the other hand, the upper limit for this resolution will be determined by the precision we want in the investigation. For example (as will be considered below), when we study a voiceless plosive of some kind, we must be able to recognize and somehow measure the length of the period of complete silence in its middle: This silence will be the same 'sound quality' as we find in pauses in the corpus. Also, we must for example be able to distinguish horisontally the parts of an affricate from the sequence of a plosive and a fricative. Such considerations points to an upper limit for the horisontal resolution considerably below the 'average phone length': A rough estimation could suggest something around, say, 10-20 milliseconds as an upper limit for horisontal resolution.

Another important factor for the horisontal resolution, a factor which will heavily influence the computational cost, concerns the time interval over which we will determine co-occurrence dependencies. Assume that we determine a time span of, say, two seconds to be the maximum interval within which sound distribution is significantly constrained. (See the last chapter for a more principled delimitation of this time span). That is, if we consider some sound quality at time x1, we assume that the distribution of sounds two seconds (or more) later will not be dependent on the sound at time x1. If we now have established a horisontal resolution of for example 10 milliseconds, this means that there will be 2000/10 = 200 horisontal positions in which we must measure the distribution. This again means that when we compute the interdependencies in an utterance, there will be at least 200 computations at <u>each</u> of the points in the utterance where we find it relevant to compute the dependency.

A final consideration as to the horisontal resolution concerns the reliability of the data. If only a poor approximation to actual speech can be achieved, such as in the below investigation where speech is simulated from text, a very fine resolution may in fact be misleading, and may imply a computational cost high above what can in fact be gained by it.

The vertical resolution is a far more complicated matter: It concerns the number of discrete sound qualities we will be able to discriminate. There is clearly a large number of possible approaches to this. Here we will only briefly touch upon the problem, since the investigation reported below utilizes a very simplified identification of sounds. Basically, the vertical resolution is a matter of the number of discrete intervals we will subdivide the audible frequency spectrum into. If we have a large number of such intervals, the number of possible combinations of these is large: That is, the number of distinct sound qualities will be larger the smaller intervals we choose for the resolution, and this number will increase very fast with decreasing interval size. The resolution must, though, be good enough at least to distinguish between elements with distinctive function, but should be kept well beyond this limit if it can be afforded computationally. The optimal subdivision of the frequency spectrum is a matter of experimental perceptual psychology.

As to the third dimension, the simplest solution will be to assign one of the values 'on' or 'off' to each of the vertical elements, determined by some critical limit for intensity value. If we have subdivided the frequency spectrum into n discrete intervals, and each of these variables can be binary-valued, the number of possible states or sound qualities will be 2 in the power of n. Similarly, if we have m discrete values in the third dimension, there will be m in the power of n possible sounds. The resolution in this dimension is, again, ultimately a matter of computational power, since the number of sounds to be observed rises exponentially with the number of values each vertical element can be assigned: The larger the number of discrete sound qualities, the larger must also the speech corpus be, if we will obtain statistically significant values. Also, the size of the database over the distributional properties of these sound qualities will rise very rapidly with increasing number of sounds, since each sound quality must be determined relative to all other sound qualities. It is therefore of interest to keep the number of possible values in the intensity dimension as low as possible. But to be able to incorporate such important features as e.g. stress, the resolution should not be below some critical limit.

The dimension of intensity seems particularly appropriate for continuously valued variation, and a possibility which immediately presents itself is to conceive these values as weighting factors in the dependency computations. A matter of considerable interest is then how the intensity value of a part of the frequency spectrum will relate to its relative impact on the dependency. This will probably have to be detemined experimentally, possibly with support from general experimental perceptual psychology. The below investigation does, though, due to the text basis for the corpus, not take intensity values into consideration at all, and the question will therefore not be pursued any further here.

A matter of considerable importance is between which measurement points and along which dimensions the dependencies are to be established. There are basically two possible solutions to this: Either we can conceive the sound stream as a string of successive single sound qualities (which amounts to a traditional symbol identification task), or we can see it as multi-layered (which is paralleled in the interrelated network approach). In the latter case, we can e.g. have a number of different sound qualities (in different parts of the frequency spectrum), corresponding to some subdivision of the frequency spectrum and intensity dimension, appearing simultaneously. If there are n such simultaneous symbols, we would thus make n observations at each horisontal position in the corpus. This would keep the number of distinct symbols to be recorded in the database low, while the number of observations would be increased. This approach finds linguistic support in autosegmental phonology and traditional feature analysis. If it can be shown that speech does in fact function according to a multi-layered model, this would drastically economize the task. Not only would the database be highly reduced, but the corpus of speech sounds could also be kept considerably smaller: Since the number of observations is n times higher than in the single-layered model, the number of observations necessary to make reliable statistical judgements could be achieved with a much smaller corpus. (Reliable, of course, on the precondition that speech is in fact multi-layered).

In contrast, the single-layered approach is far more expensive computationally. If we have n intervals vertically, and each of these can have m different intensity values, there will be m in the power of n sound qualities. This gives us a total of m in the power of 2n possible sound combinations, and these must be determined at all horisontal positions (distances). If we denote the number of horisontal positions by 'p', we will have at least

$$d = p * m^{2n}$$

number of theoretically possible records in the database over single sound qualities. Assume e.g. that p = 200, m = 2 (the simplest possible solution in the dimension of intensity) and n = 25 (the frequency spectrum is subdivided into 25 discrete intervals, a rather modest suggestion). This will give us a number of theoretically possible database records which can be written as the digit 2 followed by 17 zeros, a fairly astronomical number. Evidently, most of these will be zeros and need not be physically recorded, but the total number will probably still be so high that in order to have statistically reliable data, the corpus must be immensely large.

This may in fact be seen as evidence for the multi-layered model: In order for a child to discriminate speech sounds (and, possibly, perform a phonemic analysis over them), the amount of data needed would have to be very large indeed, and the processing of these similarly complex. The model is also well supported by the fact that phonological rules tend to operate on parts of the frequency spectrum (statable as phonological features) rather than on single phonological units.

In comparison to the single-symbol model, the multi-layered model will, with the same numbers as in the example, require only p * m * n = 200 * 2 * 25 = 10000 records in the database. The corpus needed to arrive at statistical reliability can evidently be much smaller in this model. Or, to put it differently: A much finer resolution can be afforded in the analysis of a given corpus in the latter model. This means that a continuous probabilistic grammar which is based on the multi-layered model can allow for a larger discrimination potential in the phonetic variation in speech. In psycholinguistic terms: More signification (semantically) can be coded and extracted if the grammar conceives speech not as a stream of single discrete units, but as a number of parallel streams or as a sequence of sets of unordered elements.

This difference between the two models is also reflected in the number of symbol relations appearing at each moment in speech. In the single-symbol model, there will be one dependency relation between a symbol and the following symbol, while in the multi-layered model there may be a high number of different dependency relations occurring over the same interval. This does in itself lend a higher distinctiveness to the multi-layered model (along a continuous scale), and thus a larger significational capacity.

In short: A continuous probabilistic grammar based on a scanning of a corpus of speech sounds should be feature- rather than phoneme-oriented, not only because it can be psycholinguistically and phonologically motivated, but also because it allows for a much finer resolution within a given corpus and a given practical limit of computational power.

As to the question of whether these feature bundles amounts to one or several parallel sequences of symbols and how these features interact in the grammar, the answer must probably be sought empirically.

* * *

Continuous probabilistic grammar must define a measure on syntactic dependency. This can of course be done in a number of ways: The definition of the relations between the parameters duration, frequency and intensity and how these are interrelated in the interpretation of the distribution cannot be determined without empirical testing. In particular, syntactic dependency functions can be defined with a large number of weighting factors: Dependency defined by e.g. probability of co-occurrence may be defined as varying with (i.e., weighted by) distance between the co-occurring elements, their degree of intensity, their sound frequency values etc. The co-occurrence of high-intensity sound frequencies may possibly be more salient to the syntactic function, due to their higher perceptual salience, than low-intensity frequencies etc. Similarly, it seems reasonable to assume the possibility that the lower parts of the frequency spectrum contributes more, by its higher perceptual salience, to dependency asessment than do the higher parts. If these parameters do in fact constitute real weighting factors, they must probably be sought empirically. As to their more precise values, these can theoretically be either biologically conditioned, in which case suggestions for syntactic dependency functions can possibly be sought within experimental perceptual psychology, or they can be part of the language-specific syntax, i.e., different for each language, dialect, sociolect, idiolect etc. In any case, the interdependencies between the acoustic parameters should be empirically investigated for different kinds of speech, and if a psycholinguistically realistic syntax can be established with weighting factors constant for all languages, we can assume them to be biologically conditioned, otherwise they may be language-specific.

For the present purpose, we will delimit the investigation to some simple and easily definable functions, not only because we lack sufficient data, but also because the below pilot investigation will be in a simplified form (in line with a single-symbol model) which cannot reflect the interdependencies between the various parts of the frequency spectrum and the intensities of these. The functions will make use of the concept of conditional probability and some information theoretical concepts, and we will therefore, before we continue the discussion, briefly outline the basic theoretical concepts in these approaches.

Basic concepts in probability and information theory.

The probability of an item, such as a letter or a word in printed English, expresses the chance for this item to occur in a sequence of symbols. It is expressed as a number between zero and one, and is the same as the frequency of the item (expressed as a percentage) divided with 100. If 50% of all items (in a sample) is a certain symbol, then its probability - the chance for encountering this symbol - is 0.5. For example, in printed English, if a letter has a frequency of 1 %, then its probability is 0.01.

More precisely, this is the unconditional probability. When we measure it, we take no notice of its surroundings or any other factors which could motivate its occurrence. Thus, the unconditional probability of this letter in printed English is 0.01. If, however, we assume that this is the letter 'u', and go through a text corpus and stop at each letter 'q', and then investigate the frequencies of the letters which follow 'q', we will find that the letter 'u' is heavily overrepresented. In fact, in this specific position, after the letter 'q', the chances for meeting the letter 'u' are very high, probably very close to 1.0. This means that if we take the surrounding into consideration, the probability may be very different from the overall unconditional probability. This probability is termed the conditional probability of 'u' given 'q' (here in the immediately preceding position). The conditional probability is denoted by $p(u|q)$, and for printed English we will have

$p(u|q) \approx 1.0$, while the unconditional probability may be around $p(u) = 0.01$. We will find that the conditional probability will vary with the surroundings and the conditions we specify. For example, we can compute from a corpus the probability that the word 'man' occurs after some specified sequence of words, such as e.g. 'the old'. We will find that p(man|the old) is probably higher than the unconditional probability p(man), and it is most certainly higher than p(man|read), i.e., the chances for 'man' to occur immediately after the word 'read'. The conditional probability thus expresses syntactic properties in a sequence of symbols, and a grammar of a string of symbols can be set up on basis of these conditional probabilities. This kind of grammar is often termed a 'finite state grammar' or a 'Markov grammar', and some of its properties has been investigated in e.g. Chomsky (1956) and Chomsky (1957).

Evidently, if $p(a|b) = p(a)$, then the occurrence of 'a' is not affected by the presence of 'b', and we can say that there is no interdependence between the two events 'a' and 'b'. Dependency between events is in probability theory defined by means of products of probabilities. If we, in some corpus, measure the probability of the sequence of two symbols 'ab', and these are independent of each other such that the occurrence of 'a' is not affected by 'b' and vice versa, then we will have the simple rule

$$p(ab) = p(a) * p(b)$$

i.e., the probability of the compound 'ab' equals the probability of 'a' multiplied with the probability of 'b'. If, however, the occurrences of 'a' and 'b' have some interdependence, such that $p(b|a)$ is different from $p(b)$, or $p(a|b)$ (in this case 'b' given immediately after 'a') is different from $p(a)$, then we have by definition that

$$p(ab) = p(a) * p(b|a) = p(a|b) * p(b)$$

Note that the unconditional probability of the 'a' multiplied with the conditional probability of 'b' is the same as the conditional probability of 'a' multiplied with the unconditional probability of 'b'. These laws are the theoretical definitions of the interdependence between events, but they can easily be verified by investigating the distribution in some corpus. If we know the number of 'ab'-combinations in a corpus, and we know the number of 'a' and the number of 'b', then we can easily compute the conditional probabilities from the latter formula. We can then also set up a simple expression for the degree of dependency between 'a' and 'b' by computing the difference between the unconditional and the conditional probabilities. If this difference is zero, then there is no dependency. If it is small, then we can say that the dependency is weak, and if it is large, such as in the case of the letters 'q' and 'u' in printed English, then we can say that they have a large degree of distributional interdependence.

Probabilities are often interpreted as 'the chance for something to happen' given such-and-such preconditions. Psychologically, this can be expressed as an expectancy: If we, as experienced readers of English, read a text one letter after the other and we come to a 'q', then our expectancy for the following letter to be 'u' is reflected in the high conditional probability p(u|q). If we stop at, say, an 'a', our expectancy for the following letter to be a 'u' is much lower, reflected in the low value for p(u|a). Information theory (developed in full form in Shannon and Weaver (1949)) provides an extension of these measurements of psychological expectancies by means of probabilities, although the simple probability values cannot be utilized right away.

What information theory measures, is the <u>amount of information transmitted in a code</u>. It is important to keep in mind that this does not mean the same as the everyday sense of the word. The term information is here used in a technical sense, and denotes (roughly) the capacity for a code to make semantic distinctions. (We will make this more precise in the following). Information theory is a mathematical model for quantifying structure, and does so by presupposing a paradigmatic and a syntagmatic axis for the exposition of data. Consider an alphabet which consists of only two symbols, A and B. We will use this alphabet to transmit a message about the outcome of an experiment. Let us suppose that there are a total of sixteen possible outcomes of the experiment. To make sixteen distinctions with only a two-letter alphabet, we must make a series of code words which consists of four letters in each: AAAA, AAAB, AABA ... ABBB, BBBB. We have appointed the meanings of these combinations with the receiver of the message in advance. To send the report about one result among sixteen possible ones, we must then use four letters. Now, if we have a somewhat larger alphabet, which consists of the four letters A, B, C and D, we will find that the message will also be shorter. With four letters we can make sixteen distinctions by means of only two letters in combination: AA, AB, AC, AD, BA ... CD, DD. Similarly, if we expand the alphabet to consist of sixteen letters A, B ... P, we will find that it is enough to send only one single symbol from this alphabet to report the result of our experiment.

In general, the larger the alphabet is, the shorter can the message be. In linguistic terms: The larger the paradigm, the shorter can the syntagm be. In our example, it makes sense to say that all these three messages, although the first had four symbols, the second had two, and the third had only one, <u>they all contained the same amount of information</u>. This is the content of the term 'information' as it must be understood in the present context. It concerns the number of semantic distinctions which are inherent in a sequence of symbols, and the fundamental concept is in the relationship between paradigm and syntagm. This relationship is <u>logarithmic</u>, as can be seen from our example:

$$2^4 = 4^2 = 16^1 = 16$$

The base expresses the number of symbols in the paradigm, and the exponent is the number of symbols in the syntagm. It is important to grasp this fundamental relationship between syntagms and paradigms, and it recurs in the extended definition of information value.

If we have a paradigm (an alphabet) of m symbols, then we can make m semantic distinctions with a one-letter message. If we will combine two of these letters in a message, we can take each of the m symbols as the first symbol and combine with each of the m symbols as the second of the two, which means that we can have a total of m in the power of two combinations. If we make three-letter combinations, we will find that each of the m in the power of two combinations (of two-letter combinations) can be combined with all m symbols in the third position, which means that there will be a total of m in the power of three combinations. Thus, with three-letter combinations, we can make m in the power of three semantic distinctions. This means that one such three-letter combination from the m-symbol alphabet is inherently opposed to m in the power of three alternative expressions. In general, with syntagms of p symbols from this alphabet, we can make m in the power of p semantic distinctions.

Consider another alphabet, with n symbols. We can find a number q such that m in the power of p is equal to n in the power of q. In this case, a p-letter syntagm from the alphabet of m symbols can have as many distinct forms as a q-letter syntagm from the alphabet of n symbols, and we will, as above, say that these two syntagms contain the same amount of information, since they are both opposed to equally many paradigmatic alternatives.

Now, any number can be written in logarithmic form with any positive number greater than one as base. It is customary to use 2 as base. In the present study, we will use 'log' to denote a logarithm with base 2. We can now write an equation for the two syntagms with equal amount of information as follows:

$$m^p = n^q$$
$$2^{(\log_2 m)*p} = 2^{(\log_2 n)*q}$$
$$p*\log_2 m = q*\log_2 n$$

Information theory thus defines the amount of information I in the first syntagm as p log m, and the information in the second syntagms as q log n. This equation - by the concept of 'equal amount of information' - is the basis for the definition of the technical measurement of information:

I(first syntagm) = p log m

From this we can define the entity 'amount of information per symbol': Since there are p symbols in the syntagm, the amount of information per symbol will be obtained by dividing with p, which gives:

I(per symbol) = log m

In a code, the information value per symbol equals the logarithm of the number of distinct symbols in its paradigm (alphabet). To see the fundamentally logarithmic relationship between syntagms and paradigms, we can rewrite the above expression as follows:

$$p * \log_2 m = q * \log_2 n$$

$$\frac{p}{q} = \frac{\log_2 n}{\log_2 m}$$

Thus the (size of the) paradigms are related to each other such as the inverse of the <u>logarithms</u> of the (size of the) syntagms.

In this definition of information, it has been assumed that <u>all combinations</u> (all permutations) of the symbols have been possible. I.e., when all combinatorial possibilities are utilized, the code will make use of all of its distinctive potential. In this case, all symbols will also be equally frequent (since all of them are used maximally). This means that if there are m symbols in the paradigm, each of these will have a frequency equalling 100/m%. That is, the probability of occurrence for any of these symbols will be 1/m. For a symbol A, we will have

$$p(A) = \frac{1}{m}$$

$$m = \frac{1}{p(A)}$$

We can insert this in the definition of information I = log m, and we get for a symbol A

$$I(A) = \log \frac{1}{p(A)}$$

$$I(A) = -\log p(A)$$

This is the general definition of information, by which the information value of a symbol (a discrete unit) is defined in terms of its probability of occurrence. It is essential that <u>the probability of a symbol signifies size of paradigm</u>. This is a trivial fact as long as all symbols in a paradigm are equi-probable, but it is possibly less obvious when the symbols are not equiprobable, such as is normal for natural language codes.

This may be illustrated as follows: The probability expresses the average distance between occurrences of a symbol. Assume a code with a five-symbol paradigm. If the symbols are equifrequent, a symbol A will occur on the average in every fifth position:

xxAxxxxAxxxxxAxxxAxxxxAxx

This average distance between its occurrences is identical to the number of symbols in the paradigm. Its probability (in this case 1/5 = 0.2) reflects directly the size of the paradigm. Now assume that we have another sequence of symbols where A occurs as follows:

xxAxxxxxxAxxxxxxxAxxxxxAxxxxxxAxx

From the average distance (seven symbols) between its occurrences, we will see that there is on the average space for six other symbols to be inserted between every occurrence of A. The symbol A signifies, by its frequency, a paradigm of seven symbols. The distinctive potential in A's frequency is such as it will be in a seven-symbol paradigm. It is, from this point of view, irrelevant whether there is in fact six other symbols in the code, or whether the number is higher or lower. If it is lower, the distinctive potential in the other symbols will be lower, but it is still the same for the symbol A, due to its particular probability of occurrence. This means that in a code with non-equiprobable symbols, such as is typical for linguistic codes, the symbols will have different probabilities and thus different information values.

As we said above, a probability value is not only an expression for frequency or chance of occurrence, but it is also a measure on a psychological expectancy. In our example where A occurs with a certain frequency, if we pick out randomly one symbol from this code, we will (as experienced readers of the code) have a certain expectation as to the chances for A to occur. Information value is often seen as a measure on surprise: If A occurs often, we will not be surprised to find it. If it occurs seldom, we will be much more surprised to find it if we choose a symbol randomly. The surprise is here seen as the same as the distinctive potential in the symbol, and is defined to be measurable by the information value $I(A) = -\log p(A)$.

A note on how this function behaves: The logarithm of a number between zero and one is always a negative number. When, in the definition of information value, we have a minus in front of the logarithm, the value becomes positive. This value will be zero if the probability is 1.0, and it will increase with falling probabilities. The surprise value will thus be zero if the probability is 1.0, i.e., if the event always occurs. The value will approximate infinity when the probability approximates zero, which is also well in line with the notion of surprise value: An event which hardly ever happens has a large surprise value. Information values are normally expressed in bits, which is the unit when the logarithm base is 2.

Sometimes the unit 'Hartleys' can be encountered, in which case the logarithm base is 10.

We have so far looked at the unconditional information value. But the occurrence of symbols in linguistic (and other) codes is often highly determined by context. If we return to our example of the letter 'u' following the letter 'q' in printed English, the surprise value of the 'u' in this context is evidently much lower than elsewhere. Also, if we look at the paradigm of symbol occurrences in the position immediately after 'q', we will find that it is heavily restricted compared to other positions, and the letter 'u' is strongly overrepresented in this position. The distinctive potential in this position is very low. The conditional probability of 'u' signifies a very small paradigm, which motivates the very low conditional information value

$$I(u|q) = - \log p(u|q)$$

The distribution is bound by strong sequential constraints. For the letter 'u', the information value is reduced, but for the 'non-u' letters which can be found to occur in this position in printed English (say, the letter 'a' in a name such as 'Qatar'), the conditional information value will be considerably increased compared to other positions.

Whenever there is a skewedness or constraint in distribution, such that some event is over- or underrepresented in the environment of another event, then the conditional information value will deviate from the unconditional information value. Morpheme structure conditions are typical examples of such constraints in distribution, and their strength can be measured by the conditional information value. They state restrictions on the distribution over a limited interval. Typically, at the end of the interval, the distribution will be more free. Another form of distributional structure can be found in the composition of phonemes. These are often defined as a class of more or less similar sounds. For this class, the paradigms of sound qualities which appear at, say, the beginning, the middle and the end of the phoneme interval will be governed by such constraints: After a sudden and short period of stillness, the set of sound qualities which appear (e.g. [p,t,k]-sound bursts) is very small compared to the entire set of sounds. The distributional constraints over phonemes can be recognized in another form as well: Two non-similar sounds can be classified in the same phoneme if they appear in complementary distribution. This is to say that there are strong restrictions on the environments they can appear in.

The distributional freedom is generally small within segments, and larger across segment borders. This pertains not only to phonological or morphological segments, but will be valid for syntactic phrase segments as well. The set of possible grammatical categories which can follow a determiner is e.g. smaller than the entire set of categories. Typically, we will place a phrase (a segment) border where the distributional freedom increases, and we will tend to conceive a string of syntactic units over which there are distributional constraints as some phrase segment, at some level or other.

The syntactic functions.

To define a function in a probabilistic syntax, the main concepts will be the unconditional and conditional probabilities of the occurrence of sound qualities. We concluded above that a probabilistic grammar should proceed from a multi-layered rather than a single-symbol model of speech sounds. The following discussion will, though, presuppose a conception of speech sounds as single discrete symbols without internal structure. This is partly motivated by the lack of data to test the multi-layered model against, but also by the fact that the investigation has been limited to the single-symbol representation of speech, and the framework has not allowed for the implementation of a representation according to the multi-layered model. What we have of data is a corpus of Hungarian texts, from which we can simulate (in a fairly rough manner) a stream of single-symbol speech sounds, but we can hardly impose larger variation (in the vertical dimension) on the data than is represented in the orthographic conventions. These are basically phoneme-oriented. Although a distinctive feature or even a purely acoustically oriented analysis of these phonemically oriented segments would have been possible, the resulting representation would still not contain the variability which characterizes actual speech, and the dependencies which would appear from the distribution of these 'generated' distinctive or acoustic features would basically lead us back to the phonemic inventory which we started out from.

This would, though, look somewhat differently if a corpus of speech sounds were analyzed according to the multi-layered model. We would then probably find that the close interdependency which in phonology exists between phonological units and phonological rules would emerge as clusters of interdependencies between the various parts of the frequency spectrum. It is essential that these clusters would be more or less prominent, more or less easily detectable. Thus what constitutes a discrete symbol (which either is present or absent) in the single-symbol model, will emerge to a smaller or larger extent, i.e., with a non-discrete variation, in the multi-layered model. What is important in this context is that the part of the frequency spectrum which enter into the constitution of the phoneme may be part of a larger morphological or syntactic dependency. With gradual variation, a discrete phonemic (symbol) analysis will have to record the phoneme as either present or absent, while in the acoustic reality, the gradual variation will mean that the morphological or syntactic dependency will vary continuously in a manner which cannot be reflected in the discrete symbol analysis.

Therefore, when we in the following discussion assume a single-symbol model of speech, we have drastically simplified the task, and we have - due to the preconditions for the investigation - excluded some potentially important aspects of grammatical dependencies from the definition of the syntactic functions. The definitions we arrive at may, though, well turn out to have a sufficiently general form to be applicable even in a multi-layered analysis, but this will have to be tested empirically.

We are, therefore, concerned with finding a measure on the distributional relationship between two symbols 'a' and 'b' which can reflect their syntactic binding. We define 'a' to be syntactically bound to 'b' when 'a' occurs more or less frequently in the presence of 'b' (in some distance) than it does on the average. That is, if the occurrence of 'a' is conditioned by 'b', then there is a syntactic binding between them. This can be either positive or negative: If 'a' is overrepresented in the neighbourhood of 'b', then it can be seen as positively bound, and if it is underrepresented, then it will be negatively bound. We note that the latter binding is just as much a binding as the former: Underrepresentation characterizes a dependency in occurrence just as well as overrepresentation does, although the former can be seen a matter of syntactic 'rejection' while the latter represents a syntactic 'attraction'.

The relevant measure on this binding will thus be in the form of the relationship between the unconditional and the conditional probabilities of 'a'. The unconditional probability p(a) is the overall chance for 'a' to occur. If the conditional probability p(a|b) is different from p(a), then there is a syntactic binding present. The question is now how to relate these values in a function which properly represents the syntactic relationships in a syntagm.

For the first syntactic function, we will make use of a 'pointing' function, which, in the case of overrepresentation, will point to a position <u>between</u> 'a' and 'b', and in the case of underrepresentation will point to a position outside the interval between them. The argument is as follows: If 'a' is strongly overrepresented in the presence of 'b', we will conceive 'a' as somehow 'belonging' to 'b'. An example from morphology could be the relation between a noun root and a noun suffix: The overrepresentation of the suffix in the presence of the noun (or vice versa) can be interpreted as the suffix 'belonging to' the root. Syntactically, this is represented by a common node immediately above them (or above the border between them). Similarly, if a symbol 'a' is underpresented in some position relative to 'b', we will say that it does not properly belong in this position, since it normally appears 'further away'.

On the time axis, we can assign the time value x1 to the symbol 'a' and the time value x2 to the symbol 'b'. The distance between them can be denoted the position P, defined by P = x2 - x1, i.e., the time interval between the symbols. In the 'pointing' function which we have outlined we can represent the syntactic relationship between the two symbols by a third x-value, which symbolizes the position where the relation 'properly belongs'. This third x-value can be represented by the formula:

$$x = x_2 - (x_2 - x_1) * \frac{p(a)}{p(a|b)}$$

Thus by <u>overrepresentation</u> the denominator will be larger than the numerator, and the fraction which the time distance is multiplied with will be smaller than one. This third x-value will therefore arrive closer to x2 the larger the overrepresentation is. If there is neither over- nor underrepresentation, then the fraction will have the value one, and the third x-value will equal x1. For underrepresentation (= syntactic 'rejection'), the fraction will be larger than one, and the third x-value will transgress x1 and appear outside the time interval between the symbols, further away the larger the underrepresentation is. This third x-value may therefore be seen as representing the proper position of the symbol at x1 in its particular syntactic relationship to the symbol at x2.

The overall (unconditioned) probabilities of 'a' and 'b' will normally be different, but it will always be the case that

$$\frac{p(a)}{p(a|b)} = \frac{p(b)}{p(b|a)}$$

This can be shown in the following way: The probability of a symbol is the number of occurrences of the symbol divided with the total number of symbols in a sample. The conditional probability of 'a' given 'b', i.e., p(a|b), is the number of co-occurrences 'ab' (in this position) divided with the number of occurrences of 'b'. If we the denote number of 'ab' with num(ab), the number of 'a' with num(a), the number of 'b' with num(b), and the total number of symbols (occurrences) in the sample with N, then we have

$$\frac{p(a)}{p(a|b)} = \frac{\frac{num(a)}{N}}{\frac{num(ab)}{num(b)}} = \frac{num(a) * num(b)}{num(ab) * N} = \frac{\frac{num(b)}{N}}{\frac{num(ab)}{num(a)}} = \frac{p(b)}{p(b|a)}$$

Thus, the crucial values for the syntactic binding of 'a' to 'b' is identical with the values for the binding of 'b' to 'a', which makes sense, since the syntactic binding between them is a common property to both symbols. We should also represent the binding of 'b' to 'a' with some point on the x-axis, and since the relation between them is symmetrical, this latter point will be symmetrical with the point symbolizing the binding of 'a' to 'b' about the middle of the distance P = x2 - x1. Our syntactic 'pointing' function will therefore generate the following two x-values:

$$x = x_2 - (x_2 - x_1)\,\frac{p(a)}{p(a|b)}$$

$$x = x_1 + (x_2 - x_1)\,\frac{p(b)}{p(b|a)}$$

These two points are now interpreted as representing the positions where the symbols 'a' and 'b' properly 'belong' due to the syntactic characteristics of the language. We will say that according to the distributional syntax of this language, these are the positions on the time line which 'a' and 'b' are pointing to by virtue of the syntactic properties of their distribution. We will therefore interpret the two x-values as representing points of syntactic importance on the time line.

The function will be utilized in the following way: The corpus has been scanned according to a certain vertical and horisontal resolution. Say, we have established 40 discrete symbols and chosen a horisontal resolution of 5 milliseconds. We have also chosen a horisontal range over which we expect to find significant syntactic binding values. Say, we choose 2 seconds = 2000 milliseconds as range. Then the scanning of the corpus will have to record all possible symbol combinations (which in our example will be 40 * 40 = 1600 combinations) in all relevant positions, of which there will be the range 2000 milliseconds divided with the horisontal resolution 5 milliseconds = 400 positions for co-occurrence measurements. These data are stored in a database together with the information on the total number of symbols in the corpus (which will be identical to the duration of the corpus divided with the horisontal resolution) and the number of the individual symbol occurrences, from which we can compute p(a), p(b)...p(z). On basis of these data, we can compute p(a)/p(a|b) for any combination of symbols 'a' and 'b' in any distance smaller than the established range over which syntactic dependencies are assumed to be significant. To investigate the continuous syntax of a syntagm according to this function, we will generate the two relevant points on the time line (x-axis) for all symbol combinations within the range of significance within the entire syntagm. For each symbol (which in our example has a duration of 5 milliseconds), there will be 400 relations to the left and 400 relations to the right of it, which means that each symbol will generate 800 such points on the time line, on basis of the data in the database. If the syntagm has a duration of, say, ten seconds, there will be a total of 10000/5 = 2000 discrete symbols in the syntagm, and when each of these generates 800 points, there will be a total of 2000 * 800 = 1.6 million points on the time line. (The precise number will be slightly smaller, since the number of relations for the symbols will be somewhat smaller in the beginning and the end of the syntagm). We have now defined the function to be pointing to positions of syntactic importance. This means that intervals on the time line (within the ten second's duration of the syntagm) where the density of such points is high, will be positions of syntactic significance, or positions where the 'distributionally defined grammatical density' is high.

We can therefore measure the density of these points with some statistical density function and establish a continuous scale of syntactic significance. Any such density function requires a defined range over which the density is measured. This range may e.g. be in the form of a specified number of x-values, or it can be in the form of a time interval. Given a certain range, we can set up an xy-diagram with the dimension of time along the x-axis and the density along the y-axis, and draw a curve which will show the syntax as rising and falling syntactic significance according to this function. Evidently, the smaller this range is, the more rapidly may the density rise and fall as we move along the time axis, and the fluctuations in the curve will ultimately be determined by the distributional properties of the sound qualities in the language under investigation, and cannot be predicted by any general principles. The dispersion of the points on the x-axis will not follow any predictable patterns. This means that any range (within some limits) will contain syntactic information which is not contained in any other range value, and the full syntactic analysis of the syntagm should therefore include as many range values as possible. We should therefore establish a z-axis in our diagram as well, and the syntactic structure over the syntagm will be in the form of a curving surface in three-dimensional space.

Since the range is a matter of the size of the time interval for the density computation, we will expect the fluctuations of density in some range to reflect the degree of grammatical significance at a level (phonological, morphological, syntactic) which roughly has its average segment size equalling the range. If, for example, an average phone or 'phoneme' length in the language under investigation has a duration around 100 ms, we will expect to find that the curve with a z-value (= range) of 100 ms (or, if defined in terms of number of points, in our example 800 points) will represent the fluctuations at a phonological level. If the average word length in this language is, say, 500 ms, we will expect to find the curve with z-value 500 ms to reflect word-level grammar, and so forth. These expectations are to a large extent met in the below reported pilot investigation. We refer to diagrams and discussions below, as well as the diagrams in the appendix A.

The second syntactic function: Instead of representing a syntactic relationship by a point somewhere else on the time line, possibly quite far away from the relevant symbol, we will store the syntactic information on the points 'a' and 'b'. We will now express the dependency by the inverted fraction p(a|b)/p(a). This value will be larger than one when 'a' is overrepresented in the environment of 'b', and it will be between zero and one when 'a' is underrepresented in this environment. Its limits will be plus infinity (when p(a) approaches zero) and zero (when p(a|b) approaches zero), and it will be unity when there is no syntactic binding at all. This fraction is conveniently converted into a logarithmic function, such that its limits will now be plus and minus infinity, and it will be zero when there is no syntactic binding at all. We thus define the dependency

$$\begin{aligned} \boldsymbol{D}(a|b) &= \log \frac{p(a|b)}{p(a)} \\ &= \log p(a|b) - \log p(a) \\ &= \boldsymbol{I}(a) - \boldsymbol{I}(a|b) \end{aligned}$$

which we recognize as a difference between the unconditional and the conditional information value. This can be interpreted: When 'a' is overrepresented in the environment of 'b', then it is less seldom in this environment, which means that there is less surprise in finding 'a' in this context than it is in a random position, and the dependency value D(a|b) will be positive, since I(a|b) is smaller than I(a). Similarly, if 'a' occurs more seldom in this environment than it does on the average, i.e., it corresponds to a syntactic 'rejection' between 'a' and 'b', then we will have I(a) < I(a|b), in which case D(a|b) will be negative.

For a symbol 'a', in each relation within the relevant range of syntactic significance, we can then measure the dependency value D(a|b) and add all values to a sum which characterizes a syntactic property of the symbol 'a'. This gives us a series of dependency values for the syntagm, as many values as there are symbols in the syntagm.

In the below investigation, it will turn out that the first syntactic function has a larger correlation with expected structures than this second information theoretically defined function. A part of the reason for this may be found in the high sensitivity of the latter to the relation between horisontal and vertical resolution in a single-symbol model. If there is a poor vertical resolution (as in our investigation) and a relatively fine horisontal resolution, then the discete (symbol) representation of one speech segment (a phoneme in phonological analysis) will consist in a large number of identical symbols in succession. This means that the conditional probability for a symbol to appear in its own neighbourhood will be much larger for long than for short sounds. (For example, if a normal [a]-sound is represented by 20 'a'-symbols in succession, the conditional probability p(a|a) will approximate 0.95, while the unconditional probability may be considerably below 0.1). This difference will increase by growing misproportionality between the vertical and the horisontal resolution.

A sketch of Hungarian morphology.

For the present investigation, Hungarian is ideal because of its rich morphology with an agglutinative structure and high susceptibility to morphemic segmentation. Since we are concerned with investigating a syntax with a fairly local scope, an agglutinating language is ideal for the investigation, since it presents local, word-internal syntax over the sequence of morphs, and normally has a more transparent exponency of grammatical categories than more fusional languages. Also, the syntax we are interested in establishing should be capable of pointing to the parts of speech on the time axis which have high grammatical significance, and it should indicate where, in the stream of speech sounds, the grammatical 'density' is low, i.e., where we find the borders between exponents. In the shape of a continuous curve, we will expect to find local maxima on grammatical 'loci' and local minima on the borders between grammatical constituents. The structure of Hungarian is ideal for testing this hypothesis.

We will here briefly sketch the structure of Hungarian, sufficiently for the below syntagms analyses.

Hungarian is a predominantly suffixing language. In the verbal paradigm, we find a couple of perfectivizing prefixes as well as a number of preverbal particles which are all rather loosely attached to the root, since they are readily split off from the stem when there is another candidate for the syntactically important position immediately in front of the verb stem. In the nominal paradigm, there is the sole example of the prefixed superlative morpheme 'leg-' (in addition to the variant of this, the excessive form 'legesleg-'). Except for these, all morphemic exponency is in the form of suffixation. One normally recognize only one clear example of a clitic word, the form 'is' ("also"), which seemingly can appear after any word class.

An important feature which penetrates the whole grammatical system is the vowel harmony, which exists both in the form of a front/back-distinction as well as in a less obligatory form of the rounded/unrounded distinction. The vowels /i/ and /e/ do to some extent behave neutrally, and can appear together with both back and front vowels.

NOUN stems are formed by means of a rich collection of more or less productive denominal and deverbal derivation suffixes. To these, the following inflectional categories can be suffixed, in this order: Numerus, Possessive, a possession marker ('-é') which can be independently pluralized, Case. The comparative suffix '-bb' is normally added directly to the stem, before any other inflectional suffixes.

Numerus: Singular is unmarked, plural exists in two conditioned alloforms: [k] is the unmarked form, used when no possessive suffix follows; [i] appears whenever there is a possessive suffix or the possession marker follows. Thus: 'hajó' = "boat", 'hajók' = "boats", 'hajóm' = "my boat" ('-m' is 1.p.sg. possessive suffix), but 'hajóim' (composed: hajó-i-m) = "my boats".

In its isolated form, [i] is also the sole example of a truly double exponency in Hungarian: When attached in isolation to a noun stem, it indicates both a pluralization and the presence of a possessive suffix, which in this case is the zero possessive suffix of the third person singular. ('Hajói' = "his/her boats").

Possessive suffixes: There are six basic forms, for the three grammatical persons in singular and plural, which show vowel harmonic variation. For our purposes, it is interesting to note that the singular forms share important phonological features with their plural correlates, and it is possible to analyze the plurals as some basic phonological features followed by the plural marker [k]: 1.p.sg. -m, 1p.pl. -nk; 2.p.sg. -d, 2.p.pl. -tok/tek/tök, i.e., -tVk; 3.p.sg. -Ø or -(j)V, 3.p.pl. -Øk or -(j)Vk. These features recur in the objective verbal conjugation.

The possessive marker: This is a Hungarian peculiarity which is not known from its relatives in the Finno-Ugric family of languages. It expresses that the stem to which it is attached is in the possession of something. Its form is [-é], and it is normally pluralized by [-i], such that the form [-éi] will mean that the object denoted by the stem possesses a plurality of things. It has a somewhat lexicalized variant [-ék], which always means "the family of", such as in 'Péterék', meaning "Peter's family". The possessive marker may well appear together with the possessive suffixes.

Case: There are some 20-25 case endings in Hungarian, but some of these have such a lexically restricted applicability that they may well be considered as lexicalized forms. The number of case endings is thus a matter of debate, but most authors consider something around 20 cases to be grammatically active. The most prominent forms are: accusative (-t), 3 x 3 locality cases ("inside-meaning": -ba/be, -ban/ben, -ból/ből; "upon-meaning": -ra/re, -n/on/en/ön, -ról/ről; "by/beside-meaning": -hoz/hez/höz, -nál/nél, -tól/től), dative (-nak/nek) and instrumental/sociative (-val/vel), plus some modal and time-related forms. Case endings can sometimes be difficult to distinguish clearly from postpositions, of which there also is an abundance in Hungarian. Postpositions are often only weakly stressed, and most of the case suffixes do not show any notable assimilations at the suffix border (there are only two exceptions to this, which both have a [v]-sound initially). The vowel harmony is normally invoked as criterion for determining the status of a postnominal particle: Most case suffixes obey the vowel harmony, while postpositions do not agree vowel harmonically with the stem they follow. There are, though, at least three case suffixes which do not show any change as to the vowel quality of the stem to which they are attached: the terminative (-ig), the temporal (-kor), and the causal/final (-ért).

VERBAL stems are likewise formed by a rich variety of denominal and deverbal derivative suffixes, onto which morphemes of Tense, Mode and Person (in singular and plural) are added (in this order). In addition, definiteness is expressed by the choice of conjugation: There are two sets of suffixes, one for the indefinite and one for the definite conjugation. Of dubious and well debated status is the suffix which some authors (e.g. Lavotha 1973) term a Mode inflectional suffix (potential): -hat/het. This appears before the other typically inflectional morphemes, and is always at the end of the verbal stem.

The Tense forms are present and preterite, the latter formed by [-t] or [-Vtt] added to the stem. Future is expressed periphrastically or by means of the present form.

There are three modes: Indicative, conditional and imperative. The indicative is characterized by zero, the conditional by [-nV], and the imperative by [-j]. There are some slight deviations in some forms, but these are the prevailing characteristics of the modes. The imperative shows one of the few cases of extensive assimilation processes across morpheme boundaries: if the stem ends in an alveolar fricative or the unvoiced plosive [t], the palatal glide is turned into an alveolar fricative. We note that there is no articulatory necessity in this: In other forms, we may well find the sequence [-tj] without any notable assimilation processes (although the plosive will often be somewhat palatalized). Imperative exists only in present tense, while the conditional is expressed periphrastically in the past tense.

Person: There are the two conjugations definite and indefinite. The definite form is used when there is a definite object (plus some other restricted cases) or the object is understood to be there, but not mentioned explicitly. The indefinite is used with indefinite objects or intransitive verbs. These forms are sometimes termed the subjective and the objective conjugations, because the one 'points' to the subject and the other to the object. In fact, there is a third person suffix which has the specific meaning "I ... you (sg. or pl.)", that is, it means that the subject of the sentence is the first person singular, and the object is the second person singular or plural, such as in the form 'látlak' (lát-lak), "I see you". This suggests that the conjugations, at least to some extent, express relations between the grammatical persons. Except for some slight conditioned allomorphic variation in the second person singular, the forms are very regular, and have the following characteristic sounds: 1.p.sg. indef. [-k] (but in pret. [-m]), def. [-m]; 1.p.pl. indef. [-Vnk], def. [-Vk]; 2.p.sg. indef. [-s],[-l] (but zero in imperativ), def. [-d]; 2.p.pl. [-tVk] - here the distinction between the definite and indefinite forms are expressed in the length of the theme vowel; 3.p.sg. -Ø or [-(j)V] (and some slight irregularities in a few forms); 3.p.pl. indef. [-nak], def. [-(j)Vk] (distinctions between definite and indefinite are in the preterite shown by vowel length).

What is notable with these suffix forms is the high degree of similarity with the nominal possessive forms. This is particularly true for the definite conjugation. Except for the indefinite 1.p.sg. [-k] and 3.p.pl. [-nak], we have a systematical regularity in [+nasal] expressing 1.grammatical person, an alveolar articulation expressing the 2.person, the 'loosened' approximant (as a glide or a regular vowel) articulation in the 3.person, and through the entire paradigm we find that [-k] denotes plurality, i.e., not only as the separate and unmarked plural marker on noun stems, but also as a pluralizing part of the grammatical person suffixes on nominal and verbal stems.

We also note the obvious similarity between the accusative marker [-t] and the preterite marker [-t], which in a continuous acoustically oriented grammar evidently must have very similar functions.

From our point of view, these findings are interesting by the distribution of grammatical exponents across the traditional category borders. The freedom in morpheme distribution is a trait which is a characteristic for other grammatical categories as well. For example, a large number of the case morphemes do also function as preverbal particles as well as serving as roots which can be suffixed by the grammatical person morphemes. For example, the morpheme 'be' functions as a case suffix with the meaning "into", such as in 'kéz-be' ("into the hand"). It can also function as a preverbal particle with the same meaning: 'jönni' = "to come", 'be-jönni' = "to come in". The morpheme 'ra' or 'rá' has the meaning "upon-to", in e.g. 'a ház-ra' = "upon-to the house", 'rá-nézni' = "to look upon", 'rá-m' = "upon-to me". Most of the 'case' morphemes can be suffixed by grammatical person morphemes:

'a kéz-ben' = "in the hand", 'benn-em' = "in me";
'a kéz-nek' = "the hand-DATIVE", 'nek-em' = "1.P.SG.-DATIVE";
'a kéz-ért' = "for the sake of the hand", 'ért-em' = "for the sake of me".

We can even find an overlapping in form/meaning between verbal roots and grammatical morphemes, such as in the example:

'hoz' = "to bring, carry";
'a ház-hoz' = "to (the side of the) house";
'hozzá-jönni' = "to come to (the side of) something";
'hozzám' = "to me";

Thus from the point of view of an acoustically oriented continuous grammar, the sound sequence 'h-o-z', associated with a cluster of similar meanings, can appear in almost any position in the word: as a root, as a prefix, or as a suffix. This points to a relatively free word-internal distribution of the phonetic elements which constitute morphemes.

Grammatical person morphemes can be suffixed to almost any word class root:

'kettő-nk' = "the two of us": a *numeral* (kettő) + 1.P.PL.;
'előtt-ünk' = "in front of us": a *postposition* + 1.P.PL.;
'jönn-ünk' = "our coming": an *infinitive* + 1.P.PL.;
'idézt-em' = "quoted-by-me": a *participle* + 1.P.SG. (see Tompa p.188);
'keves-ünk' = "the little of ours": an *adjective* + 1.P.PL.;
'alá-nk' = "below/underneath of us": an *adverb* + 1.P.PL.;
'mié-nk' = "ours": a *personal pronoun* (1.p.pl) + 1.P.PL.;

Most of these forms can be stressed by adding the corresponding personal pronoun in front of them, either as a prefix or as an independent word (alternatively: proclitically):

'mi-kettő-nk' = "the two of *us*";
'mi-előtt-ünk' = "in front of *us*";

Exactly the same can be done when the grammatical morpheme functions as a possessive suffix on a nominal stem:

'a barát-unk' = "our friend";
'a mi barát-unk' = "*our* friend";

This is essentially the same as we find when the morpheme functions as a verbal suffix. If the subject of the finite verb can be expressed by a personal pronoun, it is optional and will normally be omitted except when it is deliberately stressed:

'jöv-ünk' = "we come";
'mi jöv-ünk' = "*we* come";

Thus from the point of view of continuous grammar, *word classes* in Hungarian seem to be clearly fuzzy categories from a formal point of view. We note that this is a possible conclusion on basis of the rather *free distribution* of morphemes across the syntactically defined word classes. This is particularly interesting in light of the fact that Hungarian seems to be a non-configurational language which permits almost any constituent order (above word level) as grammatically possible (see in particular Kiss (1987)), although Horváth (1986) points to some notable constraints on the structure of noun phrases. This suggests a generally fairly free overall distribution of the exponents of meaning elements (constituents, morphemes).

The text corpus.

Thus, the acoustically oriented syntax of Hungarian will expectedly show a number of dependencies across traditional linguistic category borders. Therefore, to assess the influence of the distributional properties of the Hungarian speech sounds on the syntax, the investigation should proceed from the state of no predefined categories. The syntactic units, at all levels, should emerge from the investigation of the continuous syntax as a function of the distributional properties of the speech sounds. This means that we should start not only without any predefined phoneme inventory, but even without any concept of a phonemic unit or a particular phonemic level.

This requires a corpus of recorded speech sounds and the technological apparatus to analyze it with a sufficiently fine resolution. None of these has been available for the present investigation, which thus implies that it has been necessary to simulate speech from a source of written texts. As it turned out, there were no satisfying electronically stored corpora of written Hungarian available either. During my stay in Budapest in the spring 1991, I therefore collected some texts from newspapers, periodicals and printing houses in order to build a preliminary corpus for a pilot investigation. The corpus is not composed according to any principles for avoiding possible biases, but is rather selected from practical needs. Most of the collected material turned out to be crowded with typographical formatting codes which could not readily be taken out automatically, and to pick them out by hand would be too time-consuming. As it happened, the texts which I had received from the Budapest newspaper 'Pesti Hírlap' was almost entirely cleaned of formatting codes, and when I to these 2.3 Mb of newspaper text added the 0.7 Mb of literary texts (a selection from the nineteen thirties and forties, from prominent Hungarian writers) which I had received from the Lexicographical department at the Linguistic institute at the Academy of sciences in Budapest, I considered these 3 megabytes (between one and two thousand pages of text, depending on typography, or, in spoken form, approximately 83 hours of continuous speech) sufficient to give a rough representation of the phonetic surface dependencies. The presence or absence in the corpus of certain roots and the by literary style conditioned frequency of certain grammatical morphemes may have a considerable impact on a probabilistically oriented grammar based on the distribution of pre-established discrete grammatical units, but the impact from the biased selection of a continuous grammar, which investigates the distribution of sounds only, will in comparison be rather small (although not entirely absent). It is doubtful whether the below results would have been notably different if the composition of the corpus had been more in line with a principle of, say, random selection among literary sources. Of more interest, though, is the notable difference between the kind of language proficiency which underlies written language as compared to oral languge. A corpus based on written language can probably never represent the sound distribution in actual speech, and no randomly selected corpus of literary texts can ever simulate the distribution found in such special and particularly interesting speech codes as e.g. the dialects of motherese.

Since the point was to simulate actual speech, a problem of more importance was the relatively high frequency of foreign names and expressions. These occurred in both the newspaper and the literary material, but in the latter case, non-Hungarian word forms (such as Latin or French quotations, which are fairly abundant in the present material) were tagged during the coding by the Lexicographic department, and could easily be located. To save time, I decided not to go through the texts to rewrite these expressions in standard Hungarian orthography (since the morpheme structure of these words would be far off from everyday speech pattern anyhow), but deleted them automatically. This will of course produce some abrupt transitions in the simulated sound stream, but the resulting errors are probably smaller than if these phrases had been kept in the corpus. As to the large number of foreign names (with a non-Hungarian spelling) in particular in the newspaper material, there was no simple way of locating these, and they constitute, together with the non-Hungarian expressions in the 2.3 Mb newspaper text, a real error source.

Another problem of immediate importance was the very large number of numeral expressions in digit notation. I designed a somewhat simplified algorithm for turning most of these into a phonemic representation, and implemented it in a program which could handle all cardinals and ordinals up to the number 1.000.000. This was possible due to the simplissity of the Hungarian numeral system. The simplification consisted in particular in the handling of fractions and comma notation: For the former, the fraction line was ignored, such that the fraction 1/2 would be interpreted as 12. For the latter, the large number of notation conventions for numerals (in e.g. newspaper publications), such as e.g. 1 000 000, 1.000.000, 1,000,000 and 1000000 all representing the number one million, made it difficult to distinguish in a simple way between the commas of large integers and the commas of real numbers. The simplified solution consisted in deleting all commas and dots inside numerals, and maintaining as distinct only dots at the very end of a series of digits, when the following character was a non-digit. This would render e.g. 1,2 and 1.2 as 'twelve', while 12. would be interpreted as 'twelfth'. Finally, if no solution could be found for the number conversion, the program would insert a series of #-symbols (representing silence, see below) in the text. The errors for the continuous syntax resulting from these simplification will be only marginal: The main point is that some (grammatically wellformed) numeral expression appears at all: it is of less importance whether the surfacing number is one comma two or twelve. In fact, the vast number of numerals in newspaper text (dates, all kind of quantities) may give a bias to the continuous analysis, since the morpheme structures present in numeral morphemes will be fairly overrepresented compared to ordinary speech. After all, there are only a small number of morphemes (some 10-15, representing the ten fundamental digits) which recur in all forms of numeral expressions, and when the very high frequency of numerals in newspaper texts is considered, it is clear that these few morphemes will be strongly overrepresented compared to ordinary speech. For this reason, I found it convenient to delete all numerals above one million, although the cyclicity of Hungarian numerals would have made an extension to nine or twelve or fifteen digit numerals possible without any particular problems.

Numerals in Roman numeral notation were not recognized, which means that such forms as 'III', 'VII', 'IX', 'CD' etc. would be interpreted in the analysis as the sound sequences [iii], [vii], [iks], [tsd] etc. There are some of these, and they are evidently an error source in the analysis.

Another problem of considerable importance was the handling of abbreviations. There are a large number of these appearing especially in the newspaper texts. For example, in addresses, 'u.' is normally used instead of 'utca' = "street". Very frequent are the abbreviations 'kft.' and 'gfk.', which are used in business corporation names and correspond roughly to the English "ltd.". The abbreviation 'stb.' stands for 's a többi', literally "and the others", and corresponds to English "etc.", "and so forth". The most frequent abbreviations were automatically detected and replaced by their full notation, but there obviously remained a number of abbreviated forms which evidently will not conform to the overall morpheme structures of Hungarian, and these constitute a real error source.

Finally, there is a relatively small number of scattered typographical codes in the newspaper texts. To the extent that these are notated with characters which are utilized in the Hungarian orthography, they will be included in the analysis. Typographical codes in other ASCII symbols were automatically deleted.

In short, the composition and preparation of the corpus (before the redefinition) implies that there are a number of real error sources in the raw text material, although the percentage of these is probably small enough to render a corpus useful for the present analysis.

The vertical resolution.

Given this corpus of 3 million orthographical signs, the main task in the simulation of speech is to find the proper redefinition from the orthographical to a phonetically satisfying representation. Fortunately, Hungarian orthography is very close to a phonemic representation, which means that a phonemic level can be approximated fairly easy with a small number of redefinition rules. In the Hungarian phonemic system, there is a distinctive opposition between long and short of all the following phonemes:

Vowels: /i/, /y/, /e/, /ø/, /a/, /o/, /u/
Consonants: /p/, /b/, /t/, /d/, /ˆc/, /ˆd/, /k/, /g/
/f/, /v/, /s/, /z/, /ˆs/, /ˆz/, /h/
/m/, /n/, /ˆn/
/r/, /l/, /j/
("ˆ" denotes palatals, or, for the fricatives, postalveolars).

The phonemic status of the affricates in the language is a matter of debate, but many authors also set up the four affricates voiced vs. unvoiced alveolar and postalveolar, all of these also in a long and a short version. There is also a palatal (or front velar) unvoiced fricative (orthographically represented by 'h') which in some contexts has a distinctive function, but this is very restricted and the sound is in phonological analysis normally not recognized as a separate phoneme.

This system we can approximate fairly well, but from our point of view, it is not only more important to simulate the speech sounds than their underlying abstract representation: it is essential to the continuous probabilistic grammar that it is based on acoustic precategorial sensory data. But we cannot impose more variation on the material than we actually have. The orthographical representation of language is extremely reductionistic and contains only a very small part of the actual information transmitted in speech. The data we have in the form of a text corpus does therefore not allow for a study of how linguistic variation contributes to the syntactic interpretation of utterances. All we can do, is to simulate a speech sound stream with no variation, i.e., a very reduced linguistic system with only some 30-40 discrete phonetic symbols. As is well-known from experiments in speech synthesis, speech generated from such a small number of distinct sound qualities is not interpretable by native speakers of the simulated language. The representation is therefore evidently very poor.

If we look at the vowel system, we find that all vowels (as is normal) will be slightly differently articulated in the long as compared to the short variant. For all of them, it will be the case that the short vowel is more central than the long vowel. This could point to a solution where we define at least 14 different vowel qualities. If we add the audible difference in sound quality which appears when the vowels are stressed, we could expand the vowel symbol inventory to 28 discrete vowels. Evidently, the choice of any of these solutions will heavily influcence the syntactic properties of the symbols involved. If a stressed short [a] is a symbol [a1] completely different (by the discreteness of the symbols) from the unstressed short [a] as a symbol [a2], these will have completely different distributional properties: Since stress is always on the first syllable in Hungarian, the symbol [a1] will always be found in the acoustic environments shortly after word boundary, while the symbol [a2] will never be found in these characteristic environments. The representation of stressed and unstressed vowels as different symbols will therefore give more structure to the distribution than there actually is. If, on the other hand, we do not distinguish between stressed and unstressed vowels, the distributional properties of these positions will merge in the common symbol [a], and the distinction between stressed and unstressed will disappear. This means that the syntactically important prosodic information will not be represented in the continuous syntax.

Besides the important feature of stress, the vowel length is possibly important for distinguishing acoustically between the vowels as discrete symbols. In the vowel system of standard Hungarian, the vowels /e/ and /a/ show the largest acoustic differences between the long and the short variants. The long /e:/ is very high and fronted, and can in some cases be difficult to distinguish clearly from an /i/. (See also the spectrograms in Bolla (1982)). The long /a:/ is unrounded in contrast to the rounded short /å/, and it is in addition considerably more front and has a very different quality. For the rest of the vowels, the differences between the acoustic qualities of the long and the short alternants are much more modest. In the vertical dimension, where we distinguish sound qualities only, and look apart from the duration of these sound qualities, we may therefore distinguish between the following nine sound qualities: [i], [y], [e], [e:], [ø], [å], [a:], [o], [u].

In fact, from the point of view of vertical resolution, if our resolution is very broad, we may well find that the long [e:] could be collapsed with the short [i] and represented with a single symbol. The difference between these sounds would then have been in their duration only. The long [i:] is in some contexts more different from the short [i] than what is the long [e:]. Thus, from an acoustic point of view, we could have represented all occurrences of long [e:] and [i] in the corpus with a single symbol, and thus distinguished between the three high, front, unrounded vowels [e], [i] and [i:]. This will of course give very different distributional properties to the sounds: The vowel [i] will then have a frequency which is the sum of the [i] and the [e:], and we will not distinguish between the co-occurrence of, say, an [i] in the sequence [ti] and the beginning of the [e:] in [te:]. The absolute frequencies and co-occurrence frequencies would thus be very different in the two analyses.

Among the consonants, a question of immediate importance is how to handle the affricates. Although these are often interpreted as separate phonemes, and there clearly is a difference both acoustically and articulatory between for example the affricate [ts] and the sequence of the unvoiced plosive [t] followed by the fricatve [s], which thus could suggest a separate symbol representation for this affricate, the acoustic difference between the affricate and the plosive + fricative is still so small that it is doubtful whether a separate representation can be defended when the vertical resolution is as rough as it is in our context. We are mainly concerned with subdividing the set of sound qualities into subsets such that the border between the one set and the other is found where there is the largest difference in sound quality, and the difference between the members of a subset is minimal. From this point of view, the difference between the beginning of an affricate and the corresponding plosive sound is much smaller than, say, the beginning and the end of the affricate. We lose something essential in the representation if we do not add the occurrences of the beginning of [ts] to the occurrences of [t] when we calculate the frequencies. Or, rather: The vertical resolution must be very good if we shall afford to distinguish between these two [t]-sounds.

The velar nasal which appears in front of velar plosives is here considered sufficiently different from the other three nasals to be represented by a separate symbol (the character '9' is used for this). It is achieved by the simple rule 'nk/g --> 9k/g'.

The most important and serious delimitation in our data is the discreteness and phonemic status of the symbols. Speech sounds are to a large extent perceived by means of the environment they are in, and this interdependence between sound qualities across 'phone borders' is thus an important part of what constitutes the phonemic units. Also, the finer resolution we choose, the more extensive will the phonemic overlapping be: Parts of phoneme A will be acoustically identical to parts of phoneme B, and they will thus be assigned identical properties (which will be the <u>sum</u> of their frequencies) in a distributional analysis. Evidently, from the point of view of a continuous syntax, such overlapping is essential, and this is of course lost in an analysis which is based on phonemic units. A continuous syntax should rather determine sound segments on the phonemic level as clusters of syntactic dependencies between sounds.

For the redefinition, a matter of considerable importance is also how phonological rules are interpreted. In the Hungarian orthography, assimilations are represented in only a few cases. For example, a verbal root-final [t] preceded by a short vowel will in conjunction with a subsequent imperative morpheme [j] merge to a long postalveolar unvoiced frivative, and this is rendered orthographically as "ss", for example 'kötünk' (= "we bind"), but 'kössünk' (="let us bind"). Normally, though, assimilations are not represented orthographically, and we must therefore introduce these in our redefinition.

One of the main problems in this process is what boundaries we will define assimilations to cross. Assimilations apply across more boundaries in fast speech than in slow, and a regressive voicing will in Hungarian more readily function across word boundaries in fast speech than in slow speech. Similarly, the assimilations of articulatory place (palatal plosives become dental in front of dentals, dental nasals become labial in front of labials, dental plosives become palatals in front of palatals etc.) will typically apply across word-internal morpheme boundaries, but not necessarily across word boundaries in slow speech. These processes are therefore not strictly rule-bound, but are rather variations on a continuous scale. Which of these shall we introduce in the redefinition of the texts? Clearly, if we let all word-internal assimilations apply across word boundaries, we delete distinctiveness from the distribution and lose structure. On the other hand, if we retain all word boundaries, we run the risk of keeping more structure in the representation than there actually is in the oral language in general (at least for normal-speed language). A possible solution is thus to introduce assimilations across word boundaries in only a part of the corpus, and not in the rest of it, to represent a partial application of the assimilation rules. The decision on how large part should have such assimilations would, though, appear as somewhat arbitrary, since we have no data on the actual extent of these rule applications.

The solution which is adopted in most of the below analyses combines some of these redefinition problems in a common solution. In the initial stages of the investigation, it was decided not to distinguish a stressed vowel from an unstressed one, not only because the acoustic differences between them are relatively small, but also because it was of major interest to keep the number of symbols as low as possible, in order to minimize the computational cost (the analyses were initially run on a small PC) and make it easier to achieve an overview of the data. For example, the present main solution recognizes 32 distinct symbols, which gives 1024 symbol pairs in each position on the time line. If this number is expanded with the additional nine stressed vowels, the 41 symbols constitute 1681 symbol pairs in each position. The number of symbol pairs will increase exponentially with the number of symbols. For this reason, it was important to keep stressed and unstressed vowels non-distinct. The solution may, though, possibly still be defended even if distinct stressed vowels can be afforded computationally, due to the acoustic similarity between the stressed and the unstressed vowels. With the very rough vertical solution we have adopted, the difference between a stressed and an unstressed vowel is probably too small to deserve separate representations.

If, though, we ignore this difference, we remove the main perceptual cue to word boundaries: Hungarian words do always receive the main stress on the first syllable, and the stress is therefore an important word boundary signifier. To remedy for this, we can prohibit phonological rules to apply across word boundaries, since this will reinforce the word structure.

This is the solution which is adopted for the main corpus definition here, and most of the analyses will proceed from data extracted from this corpus. Some additional contrastive tests in which all stressed vowels have separate representation, or all phonological rules apply across word boundary will also be run.

The treatment of pauses is another matter of core importance.

As will become clear, the presence of silence in utterances may have a significant impact on the continuous syntax, and the reason for this is probably to be found in the fact that silence, represented by a discrete symbol, relates in a different manner to all other symbols compared to those representing speech sounds. Silence has no structure and is not related in any important systematical way to other speech sounds. (Possible restrictions on phrase-initial or phrase-final occurrence of sounds will probably have only a small impact on the distributional properties).

In the corpus, silence is represented by punctuation. Since we are concerned with simulating speech, we must somehow represent all relevant points on the time line, including the intervals of silence. If we simply omit all punctuation, this amounts to an uninterrupted flow of speech sounds for nearly 83 hours, which is of course not a particularly good approximation to actual speech. The solution adopted here for the main corpus consists in the following (somewhat arbitrarily chosen) values: full stop, question mark and exclamation mark are rendered as a sequence of ten '#'-symbols, semi-colon as seven, colon as six, dash as five, and comma as three. When in the time definition the symbol '#' receives the duration value 100 milliseconds, these punctuation marks will represent pauses of duration 1000, 700, 600, 500 and 300 milliseconds respectively.

The most frequent and for the syntax most important of these is the comma, which is mainly used clause-initial and -final as well as in paratactic constructions. These positions will also very often contain a short pause in actual speech, although they may also very often, particularly in rapid speech, be skipped or the preceding vowel may be somewhat lengthened.

All other punctuation marks (such as quotation marks, apostrophes etc.) were deleted from the text.

To sum up the vertical dimension: For the present investigation, we will primarily test the distributional properties of the a corpus defined in the following way:

Corpus A. There are 32 distinct symbols: The vowels a = [å], A = [a:], e = [e], E = [e:], as well as i, x = [y], w = [ø], o and u. For the latter, the long forms will be coded as two characters, such that 'ii' stands for [i:], while 'i' stands for [i] etc. The consonants: p,t,k,b,d,g, as well as the symbol '7' representing the unvoiced and the symbol '4' the voiced palatal plosive. In addition to the fricatives 'f' and 'v', there are the dental unvoiced '6' and the voiced 'z' as well as the unvoiced 's' and the voiced '8' postalveolars. There are further the nasals m, n, 5 (= palatal nasal) and 9 (= velar nasal). The symbols r,l and h signify the sounds they normally represent. # represents silence.

The redefinitions from orthographic to 'phonetic' representation and the 'phonological rules' are somewhat fused in the list of redefinitions presented as REDEFINITION A in the appendix B. An important feature of this corpus is that commas are rendered as a pause with duration 300 milliseconds. Phonological rules do not cross word boundaries.

In addition to the database on the distribution of these 32 symbols, another four databases where set up in order to test the influence from the biases in the main corpus. Of most importance were the impact from ignoring the distinction between stressed and unstressed vowels, the impact from commas rendered as 300 milliseconds pause, and the prevention of application of phonological rules across word boundaries. The following four additional redefinitions were therefore introduced:

Corpus B: This is identical to the previous, except that all commas are deleted, which means that the corpus simulates a speech flow in which clause- and paratactic boundaries are not signified by any pauses or slowing down of speech rate. The main purpose of this corpus is to test the influence of the frequent 'comma pauses' on the syntax, not to simulate actual speech.

Corpus C: Identical to the main corpus A, but all phonological rules apply across word boundaries. These additional rules are presented in the appendix B. Since this corpus contains no distinction between stressed and unstressed vowels, and all phonological rules cross all boundaries (except clause boundary), all structure which arises from the demarcation of word boundaries are deleted. This corpus therefore contains a minimum of structure imposed from outside.

Corpus D: In this corpus, all stressed vowels are represented by separate symbols, which adds 9 vowels to the symbol inventory. Commas are rendered as '###', and phonological rules do not apply across word boundaries. This corpus has consequently the strongest demarcation of the words.

Corpus E: Identical to the previous (corpus D), except that phonological rules cross word boundaries.

These corpora will be utilized to only a limited extent, to check the impact of the variables they represent in a few cases. All analyses in the appendix A are made from corpora A and D.

The horisontal dimension.

The duration of phones is in the present investigation primarily determined by the environment they appear in, as well as the long/short-distinction. For the vowels in corpora D and E, time did not allow to introduce the additional distinction in duration due to stress. These corpora will therefore distinguish vowels in stress position vertically only, although in actual speech they do often have a different duration as compared to the unstressed ones. (Other factors which can influence the duration of phones, such as speech rate, intonation patterns etc., were not considered).

The duration values which have been employed are all taken from Kassai (1982). She gives a thorough description of the quantity of Hungarian speech sounds, based on measurements in a collection of 900 (selected) recorded sentences with an average length of 8 syllables. Vowels are distinguished as long/short and stressed/unstressed, while consonants are distinguished by the long/short-parameter only. All speech sounds have been measured in the beginning, the middle and the end of connected intervals of speech, i.e., immediately after and before pause (except, of course, for stressed vowels at the end and long consonants which cannot appear at the beginning of utterances). These data were employed for determining the duration of the sound segments which had been defined in the corpus, but they were simplified to some extent. First of all, the distinction between utterance-initial, -internal and -final was ignored, not only because the difference in quantity between these positions is often reported as relatively small and therefore will have only a small impact on the syntax, but also because the inclusion of these distinctions would complicate the redefinition task to some extent. Since all utterance-initial and -final positions in our corpora correspond to the positions immediately adjacent to the symbol '#', and - as will be seen below - the treatment and analysis of silence and its relation to the rest of the syntagms turns out to be somewhat problematic, it seemed reasonable not to introduce an additional variable of duration in these contexts. Secondly, in the main corpus, the distinction between stressed and unstressed vowels is ignored. For these, the average value was used when the duration of both stressed and unstressed vowels was reported.

A question of some importance is how to define the environment. In the data in Kassai (1982), the length of vowels are given such as they appear between two identical consonants, such that e.g. the vowel [å] will be specified in the environments pap, bab, tat, dad etc. Similarly, the consonants are reported with the duration they have between identical vowels, such that the consonant [p] is defined in the environments ipi, i:pi:, epe, e:pe: etc. Clearly, this is an insufficient environmental definition, and it was necessary to choose either the lefthand or the righthand context as decisive. The righthand context was chosen, not only because this had been considered the most important position for Kassai's investigation, but also because phonological processes in Hungarian generally are regressive. Thus all vowels were defined relative to either a following consonant (specified for all the consonant symbols used in the redfinition), or, if followed by a vowel or silence, by their average value. Similarly, all consonants were defined relative to a following vowel, or, if followed by another consonant or silence, they were assigned the average value. The averages were computed straightforwardly from the data, without considerations to the frequency of occurrence of the combinations involved.

Vowels were specified as long and short, and were determined relative to the quality of the adjacent consonant, without regard to the duration of this consonant. Consonants are on the other hand defined as long or short and specified for adjacency to either a long or a short vowel. This slight inconsistency is due to the lack of duration specification for the consonantal environments in Kassai (1982). The list of duration definitions used for corpus A is given in appendix B. Values are in milliseconds. The sporadic apparant lefthand contexts which occur are there for purely technical reasons (the coding of long vs. short affricates) and do not indicate true lefthand contexts.

As can be seen, the time definitions are very incomplete, and although they imply a much better approximation to actual speech than some simple rule such as e.g. short = 1 mora, long = 2 morae, they are still very far from the actual durations of speech sounds.

There are also some uncertainties as to the actual values which have been used. There are some lacunae in the Kassai data, and some of the values seem to be subject to printing errors. For example, for the vowel durations which is given by [rár] = 117 ms, [ra:r] = 105 ms, at least one of them must be wrong, and the same is the case for e.g. the vowel values [jij] = 154ms, [ji:j] = 105 ms. There are a few of these, and in these cases it was necessary to estimate the values: In some cases it was assumed that the values had been exchanged, in others that there were simple typographical errors, for example that 105 ms should be read as 205 ms, and so forth.

It is also interesting to observe the considerable differences between the data given in Kassai (1982) and those found in Magdics (1969). The Kassai data seem generally to vary more than the Magdics data, and the difference between them is often of such considerable propertions as in the following case of the short vowel [ø]: Kassai stressed [bøb] = 113 ms, Magdics = 100 ms, unstressed [bøb] Kassai = 125 ms, Magdics = 90 ms. This points to the large variability of sound durations and the importance of e.g. the choice of informants to measure such durations from. It also tells us that the durations which we have utilized for the definition of our corpus is highly idealized: Natural language will never be realized with such extensive regularity.

The tables in Kassai (1982) do not contain any records on the duration of the velar nasal, and the duration values for this was therefore simply copied over from the dental nasal. This is in line with Magdics, according to whom these sounds have almost identical duration values in all contexts (although the values she reports deviates considerably from Kassai's).

Finally, since there are a lot of contexts in which the sounds are not defined, they were assigned average values for these contexts. These are plain unweighted averages. Clearly, these values will very often deviate considerably from the durations of actual speech, and do probably - together with the fact that we have constrained ourselves to the righthand context - contain the largest error source in our data on the duration of sounds.

As to the choice of horisontal resolution, I decided to use 5 ms as interval between the measurement points in the corpus. This means that the duration value of all segments would be rounded off to the nearest integer dividable with 5. A choice of a higher value, say, 10 ms, would give less variability to the duration of the segments, which is of course not desirable from our point of view. In comparison, the traditional analysis of phones into phonemic segments which are either long or short would correspond roughly to a horisontal resolution of 100 ms. On the other hand, a considerably smaller value would of course increase the variability even more, but it would also mean that the computational cost would be somewhat too high compared to what was gained.

The database.

Given a corpus redefined in a certain way, containing a limited alphabet of symbols, the duration values for these symbols, and a horisontal resolution value, the measurement for the database was done as follows: When going through the corpus, each segment was assigned a duration value according to the environment it occured in. The duration value was divided with the horisontal resolution to obtain the number of measurement points which would fall on this segment. Within the range which was considered to be syntactically significant (most of the anaylses were kept below the limit of 2 seconds interval), such sequences of symbols were concatenated in an array, and the whole range of relevant syntactic relations were counted before the sequence was moved 'leftwards' in the array, and the righthand side was filled up with a new sequence of symbols from another segment. If 2 seconds is considered to be the interval within which conditional probabilities deviate sufficiently from the unconditional ones to be syntactically significant (see discussion in chapter 3), we must make 2000 ms / 5 ms = 400 measurements for each measurement point. That is, if a segment is 100 ms, it will be represented internally in the computer as 20 symbols, and for each of these 20 times there must be 400 measurements. For each measurement, it was recorded which was the lefthand symbol, which was the righthand symbol, and the distance between them. For this combination, one occurrence was counted.

The database contains the following records: The alphabet, the total number of measurement points in the corpus, the total number of each of the symbols, and the number of co-occurrences of symbols a|b in position (distance) P. The latter records are are the most space-consuming in the database. If there are 32 symbols, there will be 1024 possible symbol combinations, and all of these must be represented for each of the 400 positions. This gives more than four hundred thousand records on symbol combinations in the database, which thus covers the syntactic relations over a span of 2 seconds of speech.

From these data in the database, we can compute the unconditional probability p(a) of the symbol 'a' by dividing the number of occurrences of 'a' (i.e., the number of times that the symbol 'a' was recorded, which will roughly be the number of times the symbol 'a' occurs in the corpus times its average duration divided with the horisontal resolution) with the total number of measurements. The conditional probability p(a|b:P) of 'a' in the distance P from the symbol 'b' will be the number of times this combination is recorded, divided with the total number of measurements of the symbol 'b'. (Again, not the number of times 'b' occurs in the redefined corpus, but the number of times it is recorded when going through the simulated speech flow and measuring the sound quality every fifth millisecond).

For a redefined corpus of approximately 3 Mb of text, the total number of measurements when the horisontal resolution is 5 ms is normally just slightly above 60 million. (This indicates that the average length of short segments, in the data we have utilized, is indeed somewhere around 100 ms). For the number of co-occurrences, there are some zeros, but there are surprisingly few of them. (These are mostly the combinations which have been ruled out by phonological rules). The distribution is of course most biased in short distances (the symbols are very close).

The density function.

In the following, we will concentrate on the first syntactic function discussed above, the 'pointing' function, defined by

$$x = x1 + P * p(b)/p(b|a)$$
$$x = x2 - P * p(a)/p(a|b)$$

When analyzing a syntagm, we must redefine it in exactly the same manner and utilize the same time definitions as was done for the corpus from which the relevant database was extracted. That is, it must be on the same form as the corpus. For example, the Hungarian expression 'az első' (= "the first") will, redefined into the same format as corpus A, be on the following form: 24 'a', followed by 13 'z', furthermore 28 'e', and 11 'l' (here there is no righthand context in the time definition data, so the average duration value 57 ms / 5 ms = 11) must be used), next 27 's', and 20 'ő'. There is thus a total of 123 measurement points in this short syntagm. Since this makes it impossible to include dependencies over more than the distance 123 positions, we can delimit the analysis to, say, 50 positions (= 250 milliseconds) in this example. This means that the first symbol, in position 1, will be related to the symbols in positions 2,3,...51, and in each of these relations, the function will generate two xvalues. If we take as an example the relation a|z:30, in which 'a' is in position 1 in the syntagm and 'z' is in position 31, we will find the following data in the database:

a|z:30	=	181601 occurrences
number of 'a'	=	4545765
number of 'z'	=	997800
N	=	60591296

(N = the total number of measurements in the corpus, the sum of all occurrences). This gives us by the formulas given above:

$$x = 1 + 30 * (4545765 * 997800 / 181601 * 60591296) = 13.37$$
$$x = 31 - 30 * (4545765 * 997800 / 181601 * 60591296) = 18.63$$

In general, if there are p relevant positions, and if the syntagm has a length of s seconds and the horisontal resolution is h, the function will generate $2 * p * (s * 1000) / h$ x-values. For example, if the syntagm is 10 seconds, and if there are 400 relevant positions and the resolution is 5 ms, we will have a total of 1.6 million x-values. In reality, the number will be slightly smaller, since the beginning and end of the syntagms will generate fewer x-values: The first position cannot be the righthand member in a symbol pair, and the symbol in, say, position 10 from the beginning can be the righthand member only in nine symbol relations.

This function was called a 'pointing' function because it points to the positions where the symbols 'properly belong' according to the distributional properties of this particular constellation. Clearly, since there may be millions of generated points in a fairly small syntagm, the values which are generated by one single pair have only a marginal impact on the syntactic structure. Also, it should be kept in mind that we are in principle dealing with a continuous syntax, which means that the horisontal resolution is theoretically approaching zero, or at least some lower limit for perceptually discriminable intervals. Thus the value p can be assumed to be, in the psycholoinguistic reality, very much higher than the 400 used in our example, and the h may be much smaller. For example, if we reduce the resolution to 1 ms, we will have 40 million points, and by 0.1 ms, the number increases to 4 billion. (As is seen, the computational cost increases very rapidly with the resolution).

The interpretation of these x-values must be in the form of a density function. Over an interval where there is a high density of such points, we will interpret this as a center of syntactic importance: Since there are more points in this area than in an equally large neighbouring area, more symbol relationships point to this area than to the neighbouring area. When we measure the fluctuations in density along the time line, we then obtain a continuous syntactic structure in which the fluctuations describe fluctuations in relative syntactic importance.

A matter of crucial importance in this respect is *to which part of the syntagm we assign the measured density value*. Assume that we go from left to right and stepwise measure the densities over the distance D. When, at some point, we get the density value y for the interval between x1 and x2, which sound shall this y-value be assigned to? Perhaps the most natural (in the sense of psycholinguistically realistic) interpretation would be that it should be assigned to the sound which is at the end of the interval, i.e, which has the x-value x2. The tests I have carried out on this question do, though, point to the sound in the *middle* of the interval as the appropriate carrier of this density. This should correspond to some delay in the perceptual processing: The assignment of density value to a sound is performed some time *after* it is perceived, as a function also of what follows the sound. This is the approach which will be adhered to in the following analyses, although it should be kept in mind that this solution is by no means the only possible one, and it may well turn out that other assignment procedures are more appropriate.

Another important question concerns *which density function to choose*. We must assume that the density of points is fluctuating on all levels, i.e., if we choose some interval I and measure its density, there will be fluctuations in density inside this interval as well. Different density functions may be more or less sensitive to such fluctuations, and one function may pay little attention to the internal fluctuations compared to another. We are in the present context mainly interested in the relative values (rather than the absolute values), the rising and falling along the time line. It is, from this point of view, not completely irrelevant which density function we choose. In the following analyses, the standard deviation for a fixed number of successive values will be used, although other functions are possible and may give somewhat different results. Some moments about the mean, the mean deviation, and the simple function x2-x1, when x2 is point number n from x1, have also been tested, and show similar, although not identical results.

This question may also be related to the previous. Since most of these density functions introduce a (weighted or unweighted) average value for the measured interval, this average - being the gravitation center or balance point of the interval - may be a good candidate for which sound to assign the y-value to. In reality, this average will probably most often be fairly close to the middle point, but not necessarily always. There has, though, not been sufficient time to test this possibility.

The analyses.

Thus the syntagms are analyzed as follows: All x-values are generated and subsequently sorted in ascending order. A range for the density function is fixed. In the first example we will use 35000 points as range. This means that we go from the left to the right of the syntagm, and for each step measure the standard deviation for the set of points constituted by the 17500 preceding points and the 17500 (minus 1) following points. We inverse the obtained value, and assign this as a y-value to the x-value at the middle of the interval. High density will give a high y-value, low density a low y-value.

The following diagram shows a syntagm analyzed in this way:

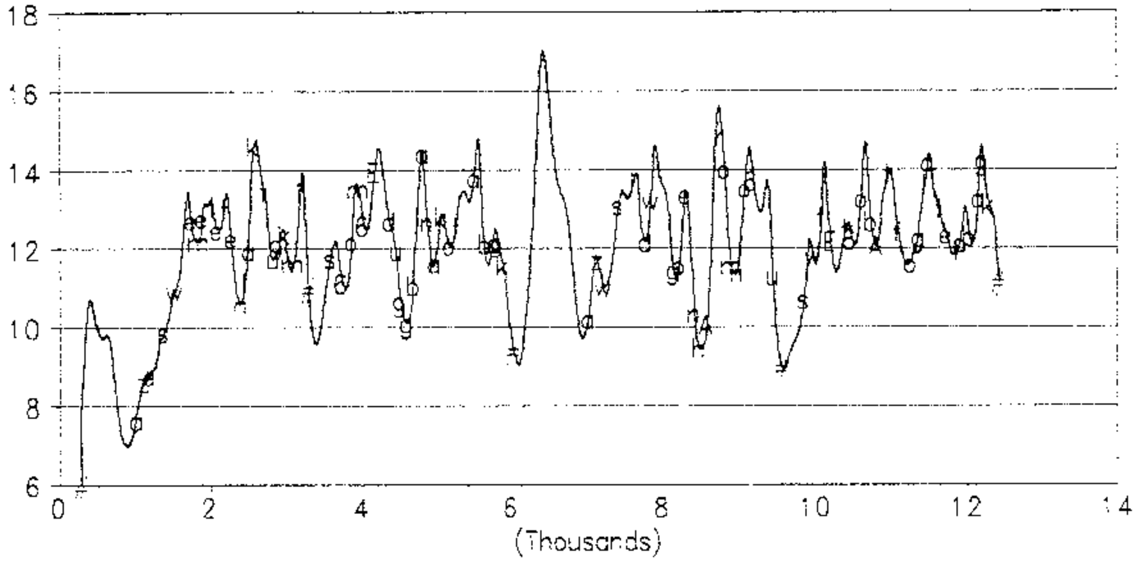


fig.1

The dependencies have been measured over 400 positions, i.e., 2 seconds intervals. The database is from corpus A. The range is 35000 points. The scaling along the x-axis is in milliseconds. The y-axis scale (the numbers on the lefthand side) is somewhat arbitrarily chosen, and it is the relationship between the different parts of the curve which is of interest in the present context. As is seen, the beginning of the curve will contain somewht deviating values, since they are based on a smaller number of positions. The y-values are comparable only for x-values larger than x = 2 seconds (and generally for x-values smaller than 2 seconds from the end of the syntagm, although this is not the case in the present curve).

The syntagm is chosen rather randomly from Sándor Márai's novel "Egy polgár vallomásai" (Budapest 1990), p.15:

'Az elsó emeleten laktunk mi, s szomszédunkban lakott a bank. Az ósidókben harom hosszú, sötét szobát foglalt el a bank, a lépcsőházból nyilt az igazgató szobája, mellette a pénztárszoba, s az udvari szobában helyezték el a könyvelést'.

The curve shows the first half of this. It can be analyzed as follows:

az	elsô	emel - et	- en	lak - t - unk	mi
the	first	lift(V)-NOM	- SUP.ESS	live-PRET-1pl	1pl
the	first	floor	- on	lived	we

s	szomszéd - unk - ban	lak - ott	a	bank.
and	neighbour - 1pl - INESS	live - PRET	ART	bank
and (the)	neighbour - our - in	lived	the	bank

Az	ős - idő - k - ben	három	hossz - ú,	sötét	szobá-t
The	ancient-time-PLUR-INESS	three	length-ADJ	dark	room-ACC
the	old times - in	three	long	dark	rooms

fog -lal - t	el	a	bank,
take-VB - PRET	PERF	the	bank
occupied		the	bank

Translated: "We lived on the first floor, and the bank occupied the flat next door. In the old times, the bank occupied three long dark rooms".

In the diagrams, the letters used to represent the sounds will be the same as in the code described above. Thus the coding of the syntagm is as follows in the diagram:

'az elsw emeleten laktu9k mi# s 6om6Edungban lakot a bank# az wsidwgben hArom ho6u# swtEt 6obAt foglalt el a bank#'

The characters are positioned on the curve at the point where the corresponding sounds start. This sound thus lasts untill the next character appears.

The curve in fig.1 shows a relatively clear example of something which can be interpreted as being close to a morphemic segmentation of the syntagm. We have defined the syntactic function in such a way that local maxima in the curve will represent points on the time line which are syntactically significant. Local minima will represent points of low grammatical importance, and we expect local maxima to point to segments or syntactically important positions in the syntagm, and local minima to point to segment borders.

If we start around x-value 2000 ms = 2 seconds (the part to the left of this is, as mentioned, based on a smaller number of points, and is not fully reliable), we see that the word 'emeleten' is divided into at least two clear wave-tops (in the following discussion, the term 'curvature' will be used to denote the part of a curve which is between two local minima): There is a local maximum approximately at the centre of the root 'emel', and another over the two suffixes '-et-en'. The root 'emel' starts slightly to left of the local minimum, and the suffix sequence '-et-en' has the final nasal on the very minimum. The following root 'lak' starts, though, immediately after this minimum, and a closer analysis of this border (i.e., with a smaller range for the density function) will show that it is situated fairly accurately between the 'n' and the 'l', i.e., on the word boundary. The next curvature contains the root 'lak' plus the preterite marker '-t', and it ends exactly where the next morpheme starts: '-unk', being the 1.person plural, occupies one single curvature, and extends from a local minimum to the next local minimum. The next peak is constituted by the personal pronoun 'mi' (= "we"), starting almost exactly at the 'm', havings its peak above the 'i', and extending into the silence symbol, which here stands for a comma in the text. In this first part of the syntagm, there is therefore a relatively clear correlation between number of prominent peaks in the curve and the number of morphemes (except for the 'double exponency' we find in '-et-en' and 'lak-t'), and the minima coincide fairly well with morpheme borders.

After the first silence symbol, the first peak appears over the single 's', which is the conjunction "and", and the following minimum is on the border to the next morpheme 'szomszéd' = '6om6Ed' = "neighbour". For this, though, the curve is not entirely convincing, since it has a notable minimum in the middle of the morpheme, while it does not have any demarcation of the following possessive suffix '-unk'. If a smaller range is chosen for the density function, a local minimum will appear at the border between 'szomszéd' and the possessive '-unk', but at this level, the local minimum in the middle of 'szom'széd' will have become even more prominent. The problem is therefore not solved by this, and this kind of problem appears rather frequently in the analyses. Clearly, such cases can be considered as counterevidence, and they do anyhow point to weaknesses at some point or other in the application of the model in the way it is done here. Except for the possible explanation as pure counterevidence, such deviations from the expected behaviour of the curve can have a number of explananations, particularly related to the possible error sources. As will be seen below, the structures do to a considerable extent depend on the definition of the corpus, and in many cases will a redefinition of the corpus representation change the curves considerably. Also, the particular choice of syntactic function as well as density function is of utmost importance, and will of course heavily influence the form of the curves.

In addition to such technical explanations and the possibility for a regular instance of counterevidence, there is also the third possibility that the curve do in fact reflect idiosyncratic properties of the segments under consideration. For example, segment borders are often downgraded in a lexicalization process. The present morpheme is not of Hungarian or Finno-ugric origin, but is borrowed from some Slavic language. The syntagms which have been analyzed during the investigation do in fact point to a tendency for domestic morphemes to conform better to the segmentation by local minima than do loan-words. This may be part of the reason here, although the data are too restricted to allow for any such conclusions, and there are many instances of the same in original Hungarian morphemes. The most important reason for the present anomaly is, though, probably that the morpheme is composed of two syllables in which the former has a back vowel and the latter a front vowel. Although it is often claimed that [e] behaves as neutral relative to the vowel harmony, the cases where it does so is relatively restricted, and its unmarked behaviour is doubtless to co-occur with front vowels. The present co-occurrence with the back vowel may therefore be the result of a statistical underrepresentation in the corpus, which causes syntactic 'rejection', and the points representing the relation will appear outside the interval between [e] and [o]. This may probably be sufficient to induce a spreading of the points in a way which creates the present structure. If this is the case, then the problem will probably not disappear with a finer vertical resolution.

The local minimum in the middle of the form 'szomszédunk' is considerably higher up than the minima at the very left and right of the word, which also suggests that the overall structure is well in line with a segmentation of the form as a word.

The following local minimum is fairly well positioned at the border between the possessive suffix and the following case suffix INESSIVE '-ban', which has a very clear single representation. Next, the root 'lak' appears again, starting only slightly to the left of the local minimum (the lateral is very short in this context: Kassai gives it no more than 37 ms of duration between short [a]'s). In this case, the root 'lak' occupies the entire curvature untill the next local minimum, and the preterite, which is now in the form '-ott' (it is not followed by any further suffixes) fills the next curvature untill the definite article 'a' starts somewhat high up. The next peak falls on this article, while the last peak before the pause falls on the morpheme 'bank', which, if we look apart from the fact that it starts slightly before the local minimum, fills one single curvature fairly well.

In short, in this part of the curve the number of peaks is exactly the same as the number of morphemes, although there is a notable anomaly in the form 'szomszédunk', which appears as segmented into 'szom-szédunk' instead of the expected 'szomszéd-unk'.

Next, the silence, which is a full stop in the text, constitutes the most prominent peak in this part of the syntagm. It is followed by the relatively clearly segmented definite article 'az'. (The definite article has the form 'a' before consonants and 'az' before vowels).

The following form 'ősidők-', coded as 'wsidwg-', presents some difficulties. It is composed of at least three morphemes: 'ős', 'idő' and PLUR '-k'. The morpheme 'idő'= "time; weather" may possibly (although there is no historical evidence for this) be further decomposable into 'id-' and '-ő'. Cp. 'ide' = "here", 'idén' = "this year", 'jövök' = "I come", 'jövő' = "future", 'a jövőre' = "next year". '-ő' is a very productive suffix whose main function is to be the exponent of present participle. Cp. also for 'idő' the possessive form 'ideje'. The analysis of this root into two parts seems therefore at least morphologically plausible, although it is more dubious from a semantic or historical point of view. In the present curve, we find that 'ős' can be relatively well segmented off: It starts precisely at a local minimum, and it ends fairly soon after the next, although its peak is not very prominent. If 'id' is a particular root morpheme, then it seems to be consituted by the next peak, although the following minimum is exactly at the beginning of the 'd'. If, furthermore, '-ő' is a separate morpheme (which, of course, it normally is, but it need not necessarily be interpreted as such here) then clearly the next and highest peak is this morpheme.

The morpheme PLURAL (realized as 'g') has no clear demarcation in this curve, and it appears only immediately before the local minimum which denotes the border to the next morpheme. This entire form is therefore difficult to segment properly with this curve, unless we allow for a segmentation into four parts, of which the former two are 'lexical' roots with separate peaks, and the latter two are 'grammatical' morphemes with one common, large peak.

The next morpheme is the front variant of the same INESSIVE case suffix which appeared around x = 5 seconds, and it is as easily segmented here as it was in the first form: It starts at the minimum (in other curves it will be found that it does in fact start exactly at the local minimum), it has the peak where the vowel 'e' starts, and it ends precisely where the last consonant ends.

From approx. 8.5 to 9 we find the very clear morpheme 'három' = "three", which starts and ends on local minima.

The adjective 'hosszú' = "long" is decomposable into the two morphemes 'hossz' = "length" and the denominal adjective formative '-ú', which is highly productive and is used in a large number of contexts. The two peaks in 'hosszú' may possibly be traceable to this fact, although the latter peak appears over the last parts of the first morpheme. Alternatively, we can see the adjective unanalyzed as a single morpheme. 'Hosszú' is in Füredeti & Kelemen (1989) recorded as the 206. most frequent lexeme in Hungarian, while the lexeme 'hossz' is number 3270. In addition, there is to the latter an alternative form 'hosszúság' (lit. "lengthness"), which is composed of the adjective 'hosszú', plus the suffix '-ság' which serves to derive nouns from adjectives. This could indicate that the sequence 'hossz' plus '-ú' is to a considerable extent lexicalized and is not readily analyzed into two morphemes. Later curves will show only one peak over the form 'hosszú', and it will probably remain difficult, in any of these analyses, to segment the adjective in an obvious way into 'hossz' plus '-ú'. This suggests that we should not stress the importance of the downward curve between the two peaks in the present analysis. We return to this form below.

We next find the silence symbol, appearing here because of the presence of a comma, due to the parataxis "long (and) dark", in the Márai text. It results in a clear separation of the two morphemes, It will recur as a difficulty later in the analysis, and we will return to it below.

The adjective 'sötét' cannot be readily subdivided into more morphemes, at least not on basis of the everyday language. In the curve, we find it split into three peaks, one above the very first sound, the next over the very center of the word (the middle consonant), and the third over the last consonant. Later analyses will show that the distinction between the first two peaks may be insubstantial, but the last seems difficult to explain, and will probably remain as a separate peak in all analyses. The reason for this is probably that it is homophonous with the accusative and the preterite. In addition, the vowel preceding it is a <u>long</u> 'e', and, as will be exemplified in the next word, there is a rule which in most cases lengthens stem-final short 'e' (and 'a)' when they are followed by a suffix, and always when this suffix is the accusative. Thus the form 'sötét' could, from a purely phonological point of view, very well have been a root 'söte' plus a suffix '-t'.

The rule is exemplified in the next form 'szobát', which is accusative of 'szoba', with a lengthening of the final 'a'. In the curve, we have a local minimum slightly after the beginning of the lengthened vowel, and a clear peak over the accusative, the '-t'. According to the hypothesis we are testing, we could have expected the middle local minimum to have been somewhat more to the right, for example in the middle of the 'á', but except for this, the two curvatures correspond fairly well with a morphemic analysis. The curving over the '-t' is rather similar to the preceding word, although it here corresponds to two morphemes, and in the former case there was only one.

The next curvature does, though, contain a number of morphemes: In fact, it contains the entire verb, including a word boundary. The verb is actually 'elfoglalt', which is decomposable into 'el', 'fog', 'lal' and 't'. In Hungarian syntax, the most salient position in the sentence is the one immediately preceding the verbal root. (See for example Kiss (198?)). If some other constituent than a preverbal particle is a candidate for this position, then preverbal particles or prefixes will normally be moved out of it to somewhere else. In the present case, the perfectivizing prefix or preverbal particle 'el' has been moved to the end of the verbal phrase to give space for the object to be immediately preceding the verbal root. If we look apart from the fact that the lefthand local minimum is situated somewhat after the beginning of the verb, and the righthand local minimum could have been slightly more to the right to conform completely to the hypothesis, the curvature comprises fairly well the finite verb in the present sentence.

As to the two last peaks in the diagram, these seem rather unproblematic: The first can be analyzed as the definite article 'a', and the last seems to contain the root 'bank' before the silence symbol, exactly as in the first half of the diagram, but this time the exponency is considerably more clear. (As is seen, the curve form over a linguistic unit is dependent on the environment).

To conclude the analysis of fig.1, there is a fairly strong correlation between curvatures and morphemes: the number of peaks is very close to the number of morphemes, and local minima tend to coincide with morpheme boundaries. In the cases of 'multiple exponency' in a curvature, the local minima still tend to fall on morpheme boundaries.

It should here be mentioned that the curve is not conforming to the hypothesis in any particular way. Compared to other syntagms which have been tested, there seems to be a fairly normal correlation between the local minima and segment borders.

Since the curve is generated from relations in 400 positions, and there are 2 x-values per position, there will be 800 x-values for every 5 ms in the syntagm. This means that the range 35000 points covers <u>on the average</u> a distance of (35000/800) * 5 = 219 ms, i.e., slightly more than 2 average phone durations. This relatively short duration may be part of the reason why such lengthy (root) morphemes as 'szomszéd' and 'sötét' are split into more than one peak. If we <u>increase</u> the range for the density computation, the fluctuations will naturally become more smooth and less sensitive to local changes, and we will expect to find larger linguistic units between local minima. Similarly, if we decrease the range, we will expect to find smaller units segmentable from the fluctuations of the curve.

We start with increasing the range, to see what happens. The following curve has all parameters identical to the previous, except for the range which is here 50000. This means that it covers on the average 313 ms or approximately three average 'phone units':

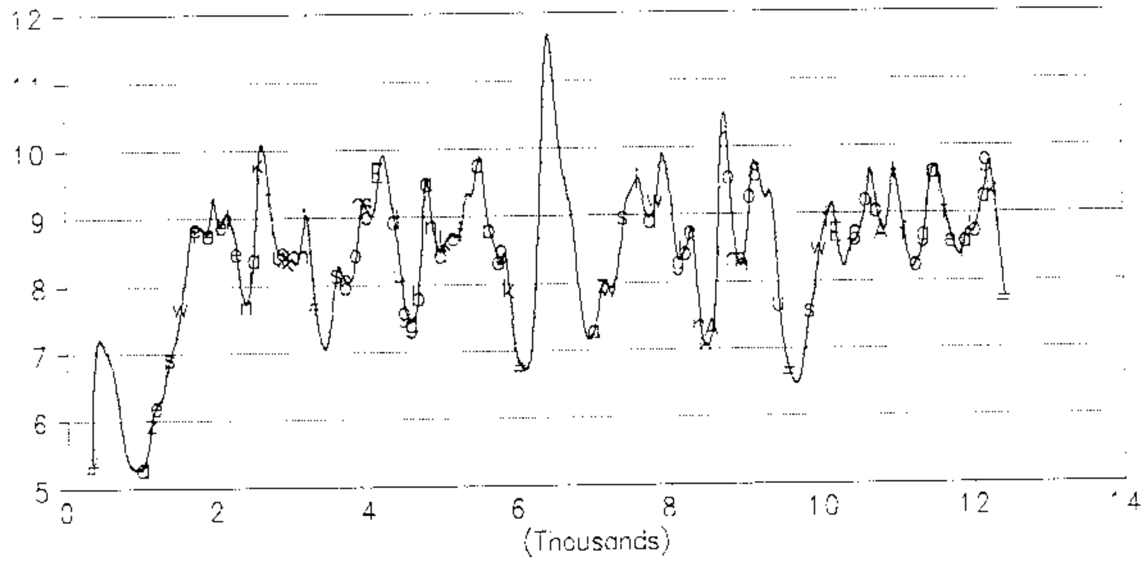


fig.2

The beginning of the curve will now be even more deformed than the previous, since the range is increased, and the incompletely filled area 2000 ms at the very beginning will be part of the density computation further to the right. The change in the curve for 'emelet' may therefore not be very reliable, although it certainly points to a significant change. The difference is, though, more obvious in the following curvature, which now covers the whole finite verb 'laktunk': The 1.P.PL. marker is no longer split off by a separate peak, but constitutes a smooth end of the verb curvature down to the first significant local minimum. The subject personal pronoun 'mi' correlates with a curvature as in the previous curve: The only change is that the relative amplitude of the curvature seems to be somewhat smaller.

Following the silence, we see that 'szomszéd' is now getting closer to a more readily segmentable unit, although it still contains a local minimum which is at least as salient as the one on the border between the conjunction and 'szomszéd'. The curvature over the INESSIVE case suffix '-ban' is still very much the same as in the previous figure.

The curve over the following finite verb is, though, fairly deformed compared to the previous curve: While both the root and the preterite morpheme correlated with separate curvatures in the 35000 curve, these are evidently on their way to be erased inthe present curve. The curve has also changed considerably over the area covered by the noun phrase 'a bank', and it seems as if the parts of the curve which are covered by the verb and the subject are about to merge into one single curvature, extending from the beginning of 'lak' untill the symbol for silence.

On the righthand side of the pause sign, there is one curvature which correlates well with the definite article 'az', as in the previous curve. The three peaks which we found in 'ôsidôk' is now reduced to two, but the local minimum between them is still not appropriately positioned, according to the hypothesized behaviour.

INESSIVE 'ben' is easily segmentable by the curve. Next, the curve form over the numeral 'három' is almost unchanged compared to the previous curve. The curvatures over the adjective 'hosszú' is, though, clearly on its way to become one single curvature, and the local minimum in its middle is no longer prominent. The same must be said about the difficult form inthe area of 'sötét', which had three peaks in the previous curve: In the present, there is only one clear peak, but the morpheme extends well into the next curvature, where the morpheme boundary is still visible as a small bump on the curve.

The densities over the noun 'szoba' plus the ACCUSATIVE '-t' are very similar to the previous curve, except that the peaks have now more equal heights. The curve to the finite verb 'foglalt el' seems virtually unaltered, and the same can be said about the final curvature over the subject 'a bank', except that the local minimum between the article and the noun is much less prominent.

To sum up this curve, parts of the curve have undergone fairly radical changes with the increase in the range, while other parts have remained almost unchanged. The tendency is, though, still to find local minima around significant syntactic boundaries. In the present curve, we find a segmentation of morphemes (INESSIVE, ACCUSATIVE), words (laktunk, hosszú) and phrases (lakott a bank, az ōsidōkben).

The middle size of these can thus be seen as characterizing the range, and we may tentatively hypothesize that if there is a correlate between a local minimum and a segment boundary, this is likely to be a word boundary. As to the morphemes, it may be noted that the most prominent of these are the case suffixes, which, as discussed above, often have a somewhat dubious status and can often be seen as clitics rather than true suffixes. In addition, they are never followed by other word-internal suffixes. This status may be part of what we see here, and if they can be conceived as separate clitics, this points to the range as a word level as well.

If we expand the range further, till the value 70000 = 438 ms, we will find that the local minima will even more suggest a phrase segmentation of the curve:

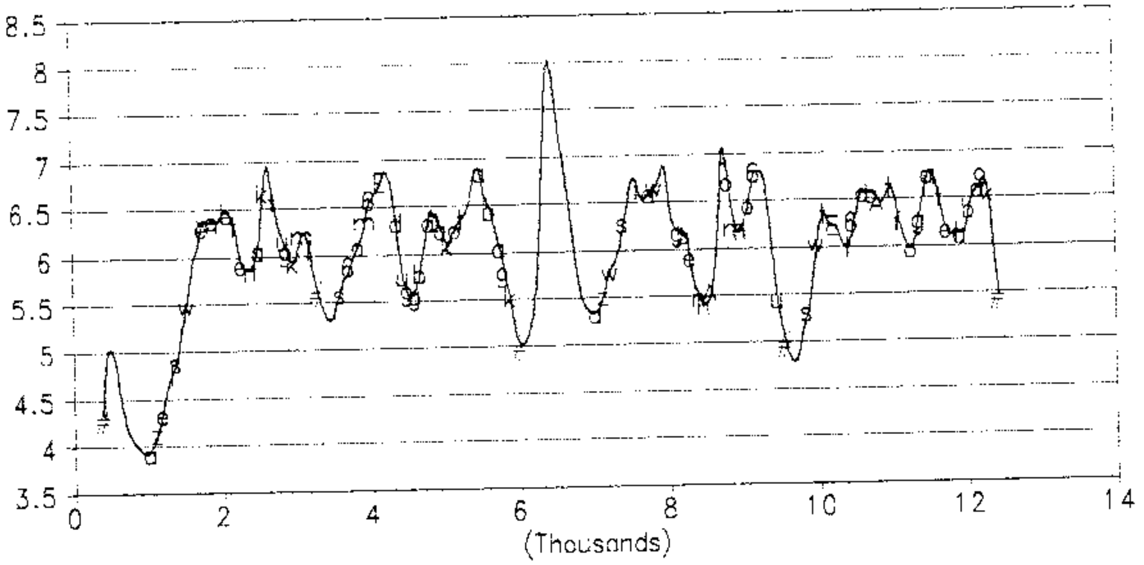


fig.3

The beginning is, as mentioned, generated from fewer points and hence should be compared with the following only with caution, but apart from this, the curve over the phrase 'az elsô emeleten' ("on the first floor") do seem to indicate a segmentation in line with the phrase boundary. The next two curvatures are approximately as in the previous curve, and cover the verb and the subject respectively. The curve over this first sentence do now indicate a segmentation which conforms very well to a phrasal segmentation.

Following the pause, we find that what constituted an isolatable curvature over the conjuction in the previous curve has now merged with the curvature corresponding to the morpheme 'szomszéd', and these now covers an area which corresponds to the conjunction + 'szomszéd' + the possessive suffix.

What follows from there untill the next pause is, though, not in line with the predictions of the hypothesis: 'lak' is now found almost at the peak of the curvature under '-ban', and the local minimum seems to indicate a border between the root and the preterite suffix. This interval, which correlated very well to linguistic boundaries in the first curve, do certainly not conform to the hypothesis that we shall find a segmentation by local minima corresponding to a phrase level at this range. What seems to be the case, is that the different parts of the syntagm need not all be properly segmented at one and the same level. Rather, what happens here is that the verb phrase, which has already been properly segmented at a lower level (range), need not be subject to such segmentation. The correlate from a binary branching analysis will be a phrase contained in another phrase: The latter will have a smaller range than the former, and both will not be 'segmented' properly at the same level. This will be seen to be the case in the below diagrams over branching structures.

After the next silence, we now find the adverb phrase 'az ôsidôkben' as one large curvature with two small twin peaks: The one evidently belongs to 'ôs', the other to 'idô', and the local minimum is now more appropriately positioned compared to the hypothesis. In the next figure, we will find that these two peaks merge into one single phrase peak. Following this phrase, we find that the words 'három' and 'hosszú' are easily segmentable from the curve. It can also be seen that there is the same direction of movement here as for the previous phrase, and the local minimum between them is about to disappear, such that they at higher levels (in larger density ranges) will apear as one single curvature.

We next find the #-symbol, which at this point can be seen to have a fairly too strong dividing impact on the curve: We would expect to find the phrase 'három hosszu sötét szobát' ("three long dark rooms") as one curvature at some level, but this seems to be impossible, and curve over this phrase have a very notable minimum by the pause sign in the middle of the phrase.

This is the problem of the impact from the commas discussed above: When commas are used to signify parataxis, it is doubtful whether the 300 ms pause is appropriate for representing the phonetic realization of such constructions. The suppressed coordinator between single words may in actual speech be signified (represented) by a short pause or a lengthening of the last sound(s) in the lefthand part, but normally not by a pause as extensive as the one found here, whch is the same as is found at the border of clauses. The reason why the comma (and thus the coded sequence ###) has been included in this syntagm, is that it conforms to the orthographical conventions, and would thus have appeared in this form if it had been in the corpus which the database is extracted from. Some alternative solutions to this problem will be presented below.

After the pause, the word 'sötét' now seems to be very close to a perfect segmentation, although it still extends somewhat into the next curvature. This is, again the same phenomenon that some units are segmented correctly at a low level, while others (not necessarily longer) such as 'hosszú' and sötét' will be segmented properly only at a much higher level.

Another, probably very important, reason why the curve seems to indicate a segmentation approximately at the beginning of this [t], may be that in the present context, where the word-final [t] is followed by a word-initial [s], the coding of this sequence becomes identical to the coding of affricates, which have a very strong internal binding between the two coded parts. Thus, the sequence [ts] will here, for technical reasons, receive the attraction which in the corpus exists between the two parts of the corresponding affricate, and this (together with the above mentioned homophony with the preterite and accusative marker) is probably the main reason why the local minimum appears on the very lefthand border of this symbol sequence. This clearly shows the impact from the poor vertical resolution: If we had coded affricates as separate sounds, the local minimum would probably have been more in line with the hypothesis in this particular case. If we had chosen a finer vertical resolution, or, for example in a multi-layered feature oriented model had otherwise distinguished the affricate from the composite, we would have expected the local minimum of the present curves to appear further to the right.

The curve part over the sequence 'szobát' is still two small peaks corresponding to the two morphemes 'szoba' and ACCUSATIVE, and the rest of the curve is also very similar to the previous.

We finally take a look at the curve with range 100000 = 625 ms:

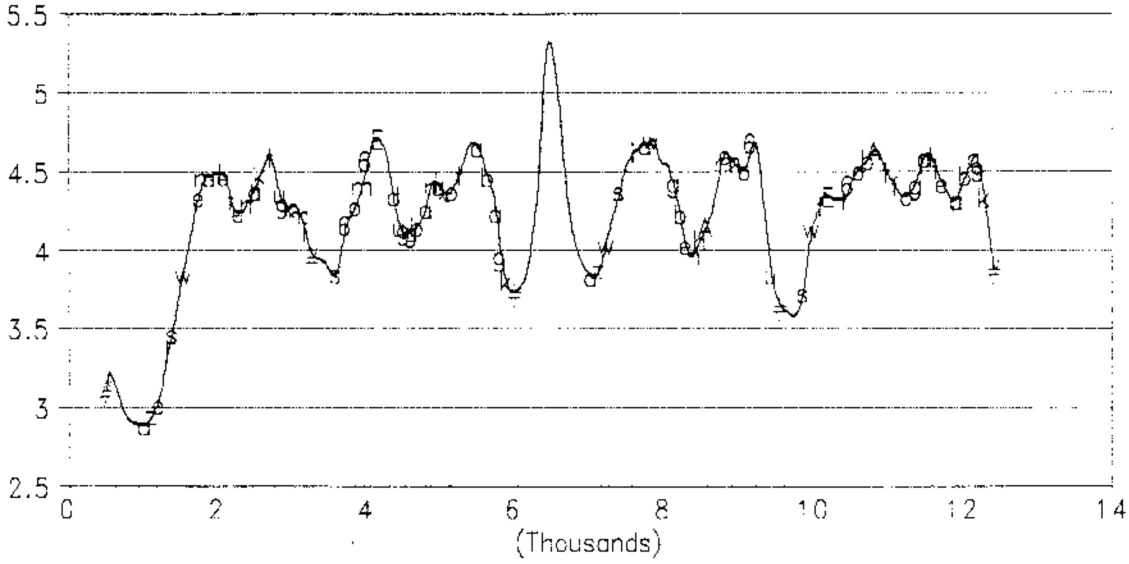


fig.4

We here find an even closer approximation to a phrase-level segmentation of the syntagm. The most notable differences compared to the previous are the following: The subject 'mi' of the first sentence do now form one segmentable unit together with the preceding verb (i.e., the local minimum between the verb and the subject is about to disappear). The local minimum after (or on) the pause has moved closer to the conjunction. The curve over the sentence between the pauses is divided into three peaks, and the finite verb starts on the top of the middle peak. If the range is increased even more, up to 150000, we will find that this sentence is divided into two peaks, one for the conjunction + adverb 's szomszédunkban', the other for the finite verb + subject 'lakott a bank'. The local minimum between these two curvatures arrives fairly close to (but not exactly on: see the below diagrams) the border between the adverb and the verb:

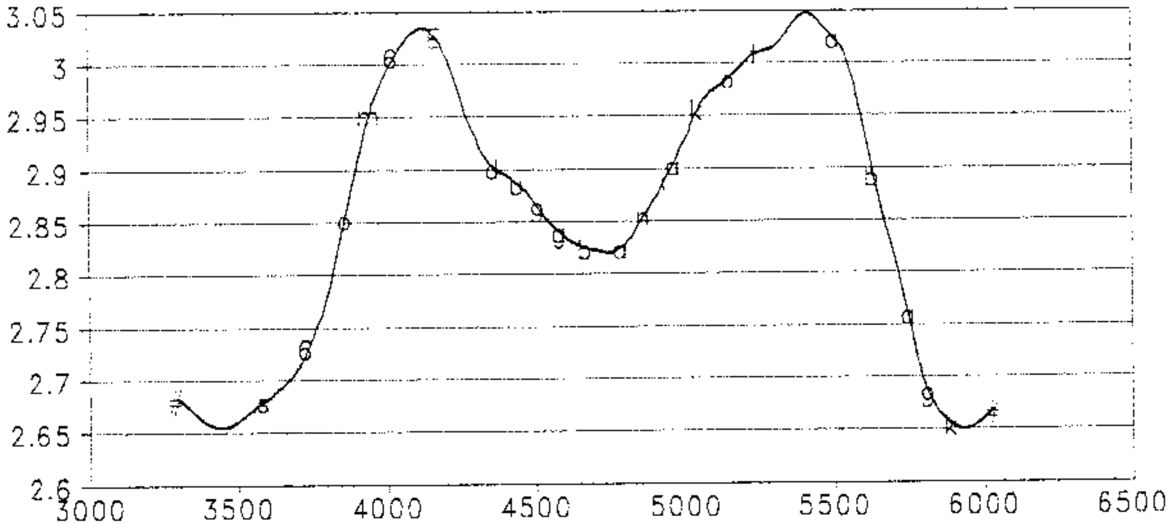


fig.5

'Az ôsidôkben' (in fig.4) can now be segmented as a complete phrase constituent. The numeral ('három') plus the adjective ('hosszú') is close to comprising one single curvature. After the (obviously misplaced) comma-silence we have 'sötét szobát' as two peaks, but very close to one constituent. And, most importantly, the finite verb phrase ('foglalt el') is at this level approaching a much better segmentation. Finally, the subject noun phrase ('a bank') comprises one single constituent.

The general tendency seems to be that most linguistic boundaries coincide with a local minimum at some level or other. (This conclusion is based not only on the results from the present syntagm, but also on the experiences from the investigation of a number of other syntagms). Also, the number of local minima which do not have a boundary correlate has in the present syntagm been small, and this is also the general tendency.

In the present syntagm, the more we expand the range over which the density of generated x-values is computed, the more we seem to approach a phrase-level segmentation of the syntagm. As we moved from the range 35 to 100-150 thousand, we passed from a basically morphemic to a phrase-level segmentation. Not all boundaries have been correctly represented at all levels, but it has been the case that most boundaries in the syntagm have had a localminimum correlate at some level. The syntax is - per definition - not discrete, which means that it does not have to presuppose discrete levels either, and there is no need to postulate the segmentation of what will be recognized in discrete grammar as a particular kind of unit at a particular level.

There were some instances of single curvatures covering two morphemes. If we go in the other direction, and decrease the range, we find that these cases of 'multiple exponency' in the first curve (35000) will be relatively well segmented at lower levels. The following diagram is an exerpt from the beginning of the 10000 range curve, where we find the two cases of 'double exponency' from the first curve: '-et-en' = DER.SUFF + SUPERESSIVE, and 'lak-t-' = 'lak' + PRETERITE:

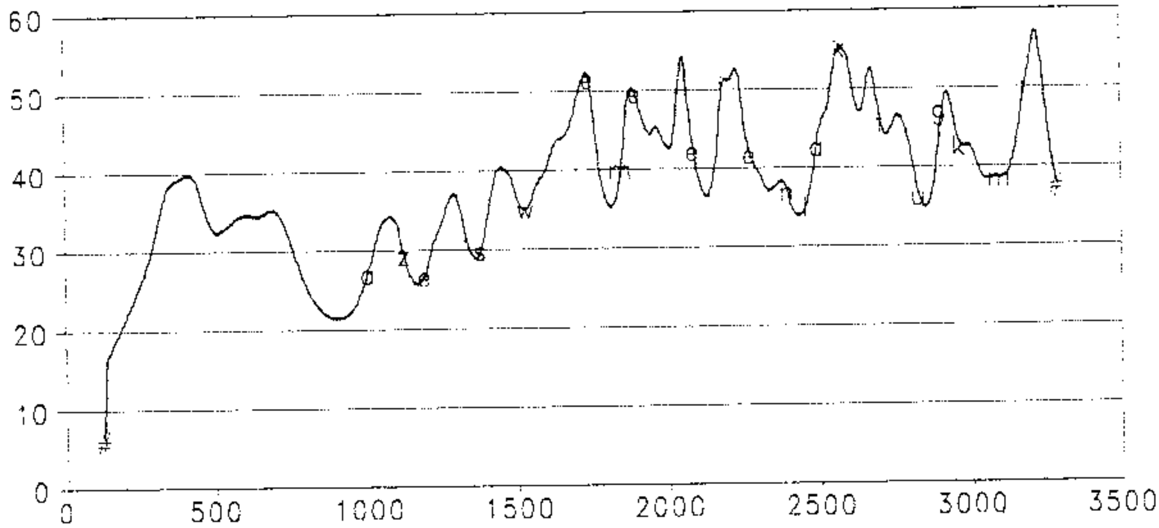


fig.6

Here we can see that the superessive '-en' (the suffix is strictly speaking only '-n': the vowel is a theme vowel which appears after consonant-final stems only) appears as a rather weak bump on the end of the curvature. The same is the case for the preterite, and here we find that there also is a peak immediately in front of it, i.e., a peak which cannot be interpreted as representing any particular meaning element in itself. Clearly, if we continue downwards into smaller ranges, the peaks become more and more numerous and dense, and lower-level units can be defined to exist. The present function do, though, not seem to be appropriate for the study of these, in particular not with the phonemically based definition of the symbols which we have here, in which there is no internal structure in the symbols.

The ranges we have looked at amount to different views on the same basic distribution of points along the x-axis, and they are all simultaneously present in the syntagm. They constitute a separate continuous scale in the syntax, along which we find variation in the appearances of local maxima and minima. We can thus represent the continuous syntax along three dimensions:

1) the time dimension along the x-axis,
2) the density dimension along the y-axis, and
3) the range dimension along the z-axis.

The continuous syntactic structure over a syntagm will thus be a curving surface in three-dimensional space. The following figure indicates (in a rather poor fashion) the nature of this syntactic structure. It is generated from another syntagm, but can serve the purpose of showing the basic features of the surface.

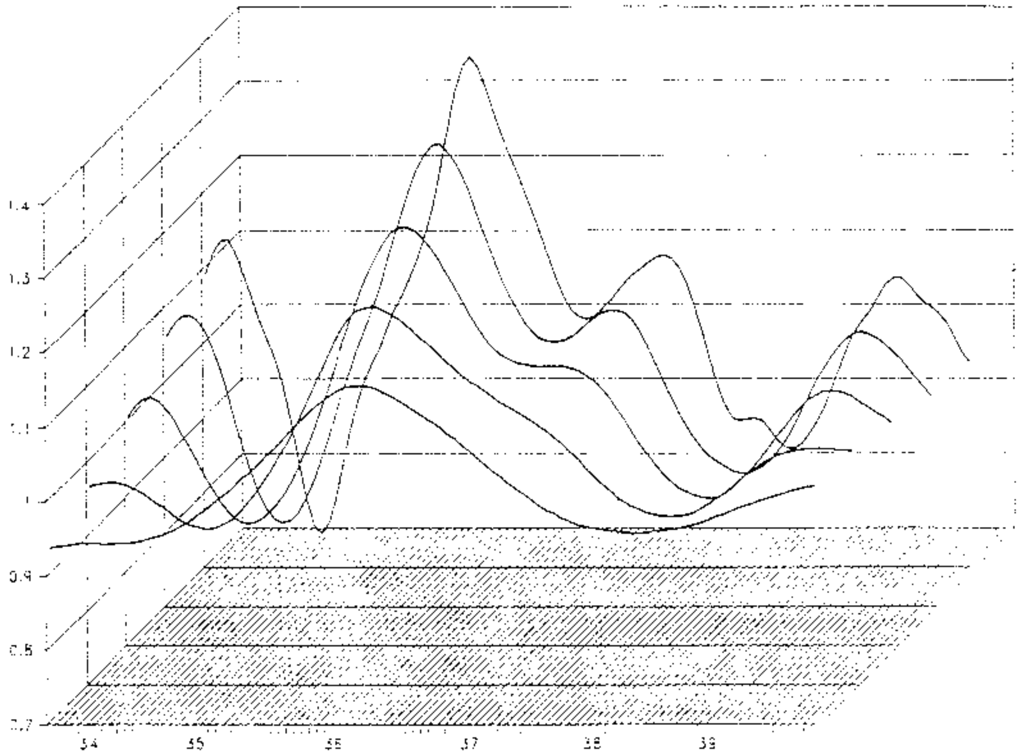


fig.7

The z-axis is here the axis moving 'towards us'. The axis has its lowest values at the 'back wall', and has increasing z-values as we move outwards. The curves are of the same kind as the above, and can be seen as randomly chosen from the surface, which is continuous. The basic purpose is to illustrate the nature of this surface: With decreasing z-values, i.e., as we move towards the back part of the surface, the number of local maxima and minima will increase, and as we approach z = 0, this number of local curvatures will (theoretically) approach infinity. In the other end of the surface (close to us), the local curvatures will flatten out and approach a straight line (i.e., no variation in density) as the z-value approaches infinity. The syntactic structure proper is to be found in some local area, such as the z-values we have studied above. In this area, the important property is that since there are more local maxima and minima for low z-values than for high, the curvatures must be branching. As is seen on the present illustration, the one local maximum on the main curvature in the front is bifurcating into three local maxima in the back part. As can be seen, there is also a beginning bifurcation on the lefthand side of the back curve. These bifurcations of the curve will thus - according to the hypothesis - point to syntactic segments or positions, and as we have seen in the analysis of the present syntagm, the local minima tend to coincide with linguistically relevant boundaries. This means that the bifurcations on the surface will correspond roughly to the subdivisions into smaller segments at different levels of syntactic analysis.

Evidently, the whole surface contains relevant syntactic information, but we can extract the most important information by taking out the local maxima only from a set of such parallel xy-curves. Since we are not interested (in the present context) in the exact y-values, we can represent these by the x-value and the z-value only, and plot these in a plane. We will then obtain a <u>diagram of the bifurcations</u> on the surface. The following figure shows the result of such an analysis of the first part of the above syntagm:

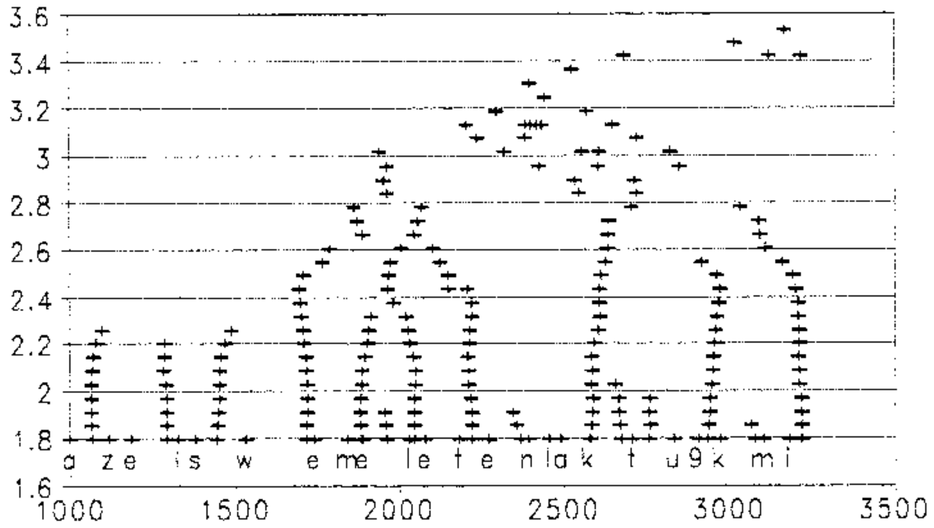


fig.8

Here , the y-axis represents the <u>z-values</u> in a three-dimensional diagram. The scale along this y-axis is here logarithmic: The range values have deen increased exponentially when going from the low to the higher z-values, simply in order to make it easier to overview the structure. The y-values in the diagram is given by y = log10(z * resolution / (number of positions * 2)). This means that if one computes 10 in the power of the y-value, one gets the average interval (in milliseconds) over which the density computation ranges. Thus, for orientation, z = 10.000 gives y = 1.8 (the lowest series in the diagram), and an average time interval of roughly 63 msec, z = 35.000 gives y = 2.34 and time interval 219 msec, z = 70.000 gives y = 2.64 and time interval 437 msec,, and z = 150.000 gives y = 2.97, interval = 933 msec. The computations for the present diagram covers the range from z = 10.000 to z = 2.000.000.

The characteristics of the above curves can be recognized in the present diagram. For example, in the first sentence, the curvatures over the two morphemes (at about x = 2.3 seconds) 'lak' and 't' merges at a rather low level (around y = 2.0, corresponding to z = 16.000), while for example the bifurcation which splits the curve over the initial phrase 'az elsö emeleten' is situated rather high up, around y = 2.8.

As is seen, this representation is in the form of a branching structure, and the branches correspond roughly to linguistic segments at various levels. In fact, the present diagram is fairly close to a traditional binary branching syntactic tree. This should not be too surprising: If all local minima at all levels coincide with linguistic boundaries, then the branching structure will be identical to a binary branching syntactic tree. Bifurcations in the branching structure will correspond to branching nodes in the syntax. If, therefore, a proper redefinition in the corpus as well as an appropriate syntactic and density function can be found such that the match between local minima and linguistic boundaries become more perfect than we have found it to be in the above analyses, then we can in fact generate a syntactic tree by means of the distributional properties of the speech sounds in an utterance.

This is indeed a tempting possibility, since it would imply that the syntactic structures of utterances are immediately accessibly in the very phonetic form. The rest of this study will concentrate on this possibility, and we will see how close we can get to a syntactic interpretation from the information in these xz-diagrams.

In the structure in fig.8, which contains much higher z-values than the above curves, also shows that the entire sentence is comprised in one single curvature when the z-value (the range for the density computation) becomes sufficiently high: In the present diagram, this curvature seems to have its bifurcation point around 3.3. From this node, there is a branching to the adverbial on the lefthand side and subject on the righthand side. The node to which the verb 'laktunk' is attached is somewhat more difficult to determine, but it seems most reasonable to interpret it as attached to the branch extending to the adverbial. This emerges as a clearly distinguishable constituent over the area x = 1000 - 2400. The verb 'laktunk' extends from 2400 to 3000, and the subject 'mi' is to the very right. The tense morpheme 't' joins the root at a very low level. The branch from the 1.P.PL. '-unk' at 2900 bends leftwards towards the verb root, and we can hypothesize a common node for the entire verb around x = 2700, y = 2.8. (An alternative interpretation, which is partly supported by the below data, suggests that the person marker does rather branch together with the subject 'mi'. This is also motivated by the common referee for these morphemes).

A tentative binary branching analysis of the diagram could thus look as follows:

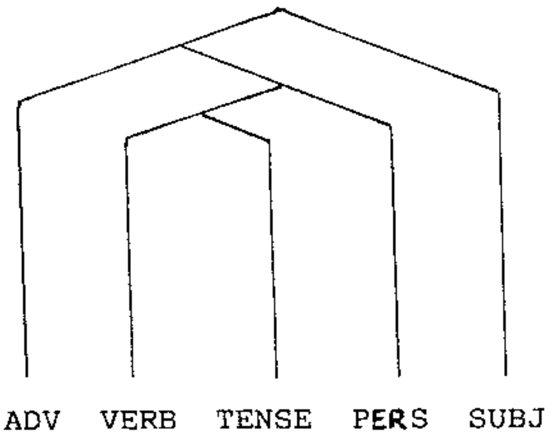


fig.9

The next sentence, which is coordinated (by the conjunction 's') with the previous, appears as follows when local maxima are extracted from the syntactic structure:

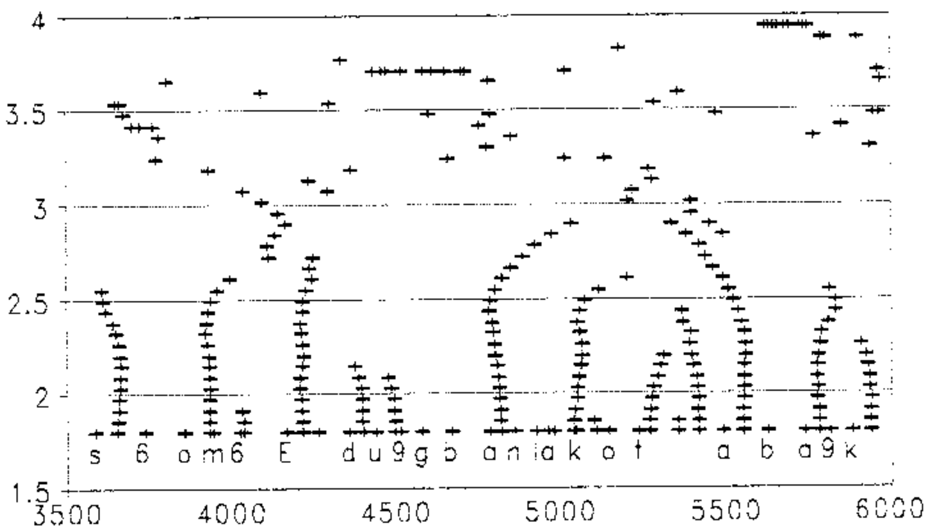


fig.10

The conjunction seems to be branching leftwards (or possibly vertically), which corresponds well to what we could expect. The entire sentence structure is not as clear as the previous (partly because an insufficient number of xy-curves has been scanned for local maxima), but it seems possible to assume a sentence node approximately in the centre, at about y = 3.3. Again we find a branching to the adverbial ('szomszédunkban') to the left, but now the verb seems to be attached to a common node with the subject around x = 5250, y = 3.1/3.2. The number of points in this area is, though, actually to small to make any conclusions on this.

There are two main features of this structure which relatively clearly separates it from a normal branching analysis: The most important is the rightbranching of the suffix '-ban' around x

= 4800. We recognize this from the curves above: Even for very high z-valaues, the local minimum which should be interpreted as a prominent boundary was of the lefthand side of this suffix, and this is what is reflected here. The second anomaly is the very clear branching from the definite article 'a' slightly to the right of x = 5500. In a binary branching tree, this should be attached to the following noun 'bank' at a low level, but here it seems to be attached to the verb independently of the noun. These two anomalies are probably at least partly determined by the error sources in the investigation, and the noun phrase 'a bank' will be found to appear in a more proper form in the following sentence. Compare also with the structure in fig.18 below.

Although the definite article stands out as somewhat too prominent, it still seems reasonable to interpret the noun phrase as constituting the subject as a unit, joining the verb probably in a node around y = 3.

The next diagram shows the second half of the syntagm we are investigating:

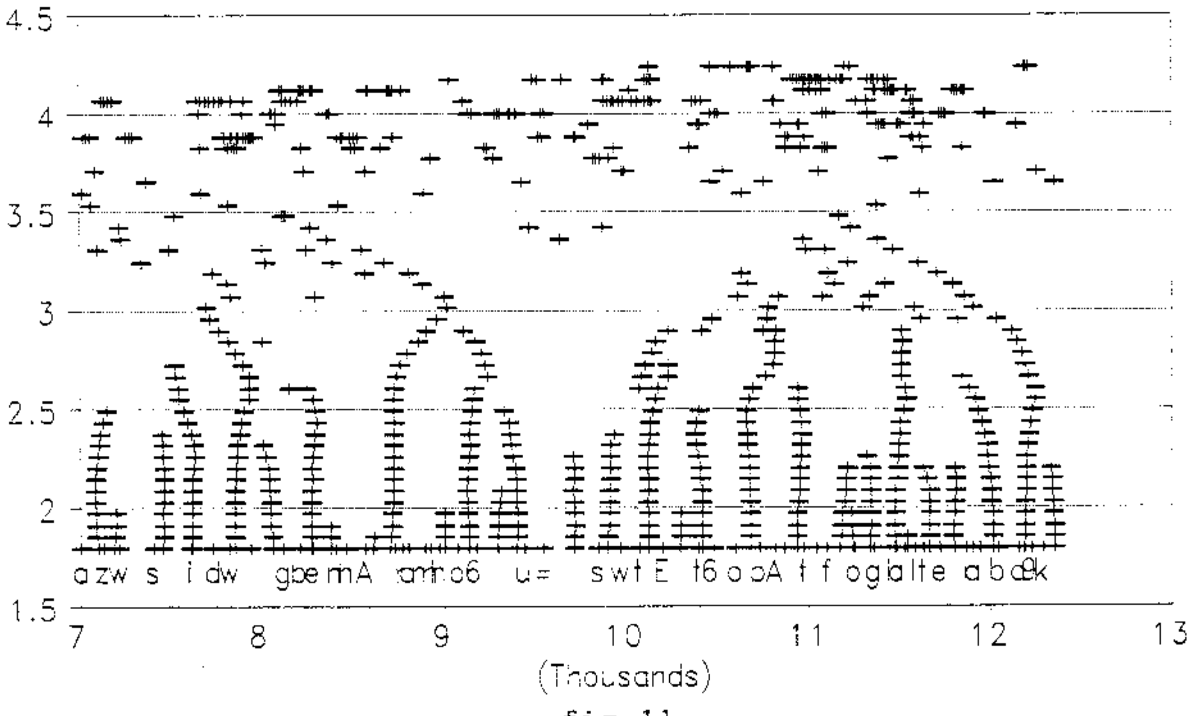


fig.11

The structure is clearly divided into two halves, with the major division in the middle, just above the pause sign, although this division is in the middle of the object noun phrase. The orthographical convention which requires commas in such paratactic constructions imposes a boundary into the phrase which is equivalent to a clause boundary. To show the extent of the impact from this pause symbol, defined as 300 ms silence, on the the syntactic structure, the following diagram is generated from a syntagm identical to the previous except that this comma is removed. The database is otherwise the same as for the previous syntagm:

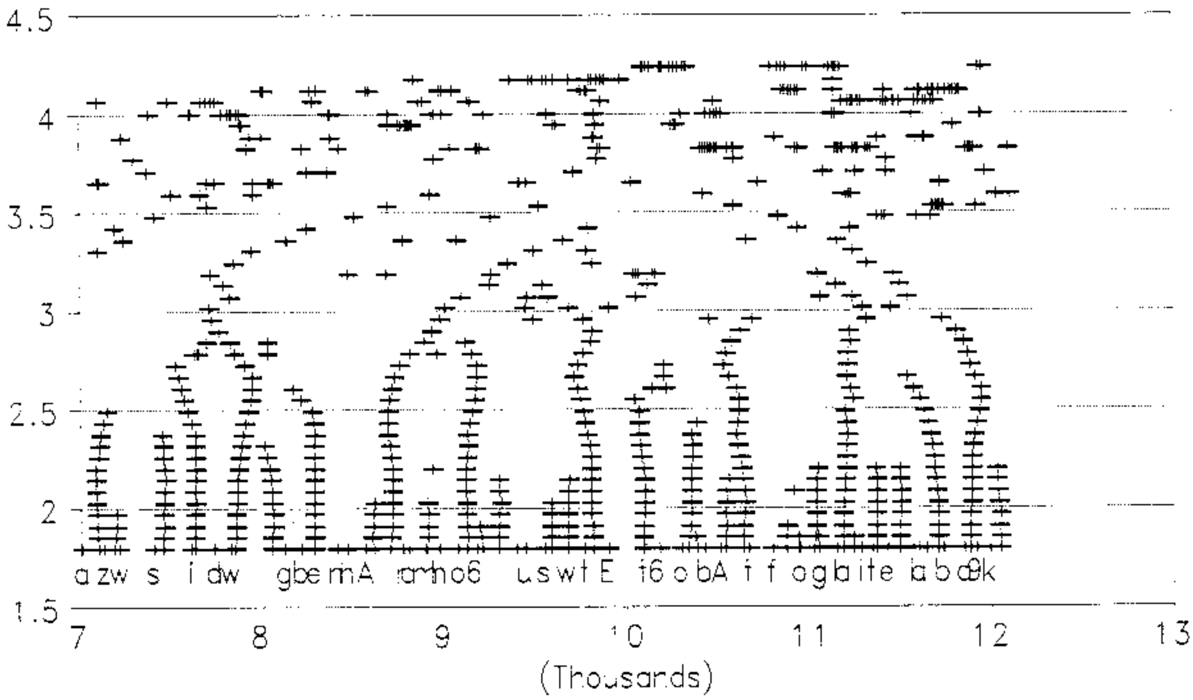


fig.12

The clear division disappears when the comma is removed. On the contrary, the whole object noun phrase now constitutes one unit, and 'három hosszú' now branches into what seems to be a common node around x = 9.8, y = 3.4. Before this phrase, we find the adverbial 'az ôsidôkben' as a separate constituent outside the main cluster, and it is possibly to interpret it as branching into a sentence node around x = 9.8, y = 3.7/3.8. The verb phrase 'foglalt el' from 10.8 to 11.6 is represented by a single branch, comprising the root as well as the TENSE '-t' and the ASPECT 'el', and the whole phrase seems to be adjoined to the following subject in a node around x = 11.5, y = 3.1. The slightly leftbranching line from the definite article 'a' towards the verb looks in this diagram very similar to the previous diagram, but as will be seen below, when some of the redefinition conditions are altered, the structure improves considerably.

To abstract a tentative binary branching constituent structure from this diagram, the main problem seems to be in the object, in which there are several possible solutions. The following discrete analysis presents one solution, with dotted lines to indicate branches of unclear status:

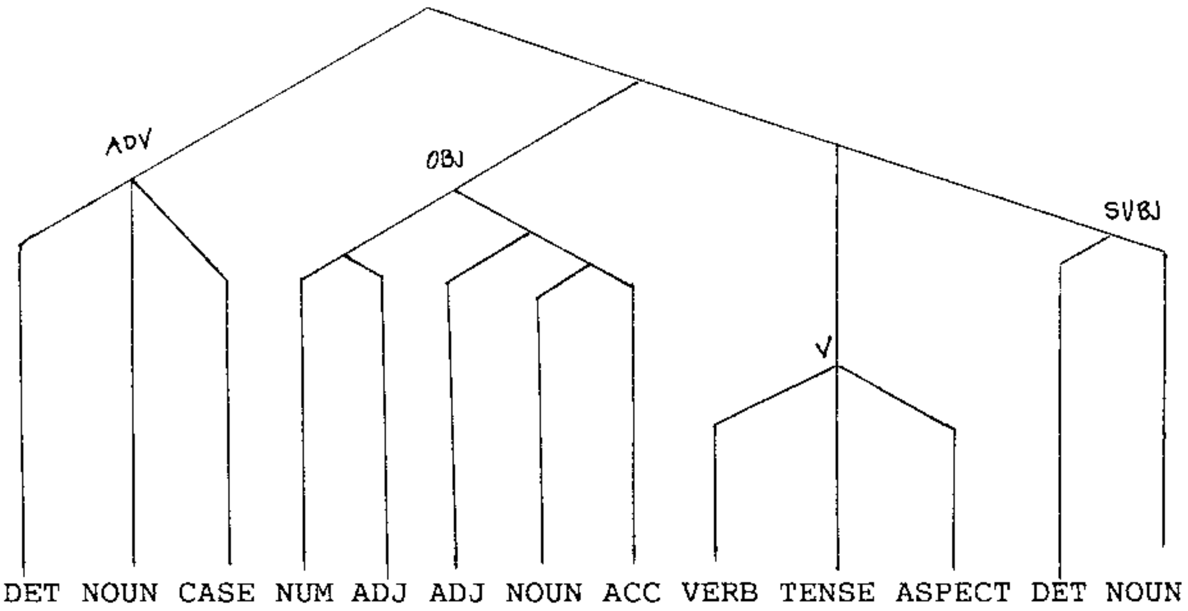

ADV
OBJ
V
SUBJ
DET NOUN CASE NUM ADJ ADJ NOUN ACC VERB TENSE ASPECT DET NOUN

fig.13

In the adverb, there is a plural marker immediately in front of the inessive suffix, but this has been suppressed here since it does not appear very vlearly in the structure. It will emerge more clear below. The adjectives in the object noun phrase may be seen as branching either way.

To illustrate furthermore the importance of the corpus definitions and the impact from the punctuaton rendered as silence, the following structure is made by means of a database over the distribution in corpus B, in which all commas have been deleted. This is obviously a strained representation: It amounts to a stream of speech which is completely uninterrupted except at full stops, question marks etc. This representation is therefore at least as exaggerated as the inclusion of the comma, but it has been made to test the impact from the comma. The syntagm has now been coded entirely without commas (to have it on the same form as the database), and the structure is as follows:

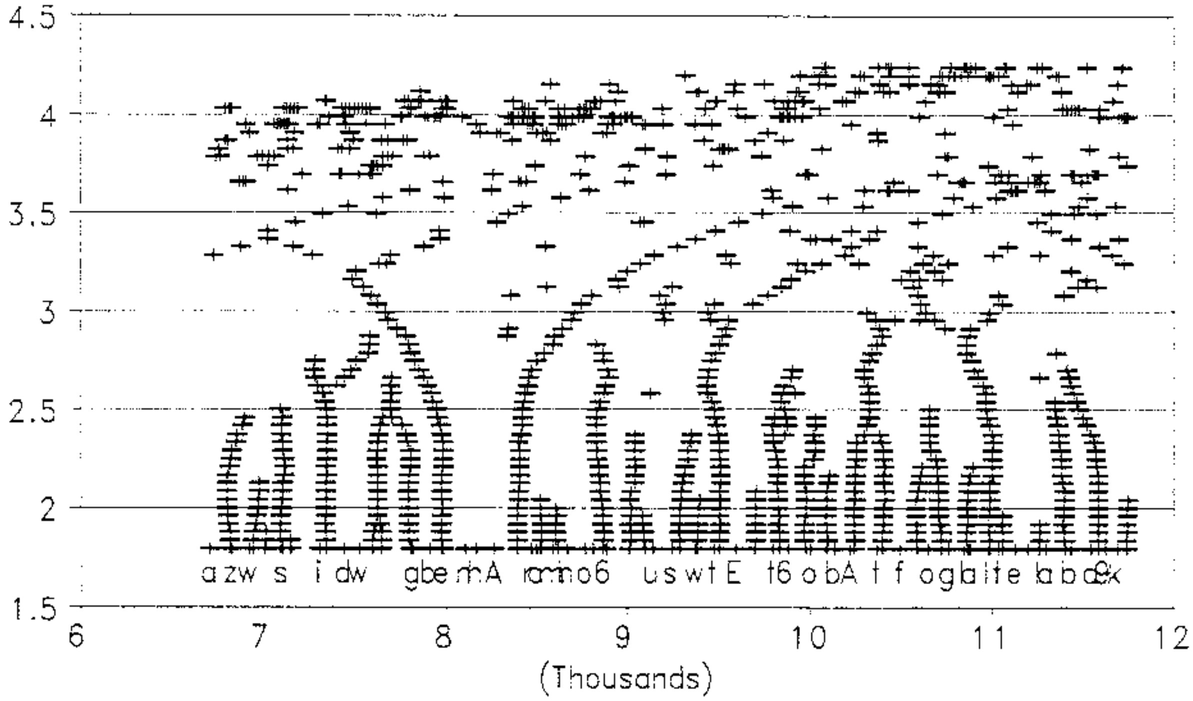


fig.14

There are some very notable differences compared to the following: First of all, the subject is no longer clearly attched to the rest of the syntagm. This must probably be due to the fact that the pause sign after it has been removed, to the effect that it is immediately followed by another constituent which draws it rightwards. It should, though, be noted that it is still inclining leftwards towards the constituents to which it belongs syntactically. The initial adverb has also been notably altered: The affix INESSIVE does now seem to dominate the whole phrase, the plural marker seems to have got (more clearly) a separate branch, and there are some other minor changes. In the object, the noun (or more exactly the latter half of it) is more directly attached to the verb rather than to the rest of the object phrase, while the former half of the noun seems to constitute a unit together with the latter half of the preceding adjective. (As mentioned above, this is probably due to the impact from the coding of the sequence t+s identical to the similar affricate). Except for these anomalies, the verb phrase has a more clearly rightbranching structure than the previous, and a somewhat remarkable feature is the clear segmentability of the constituents in spite of the artifical conditions which has been introduced in the corpus definitions.

The following diagram shows the structure generated over the same syntagm on basis of the data extracted from corpus D, in which all stressed vowels are represented with separate symbols and phonological rules do not apply across word boundary. This corpus thus contains the strongest word boundary demarcation criteria:

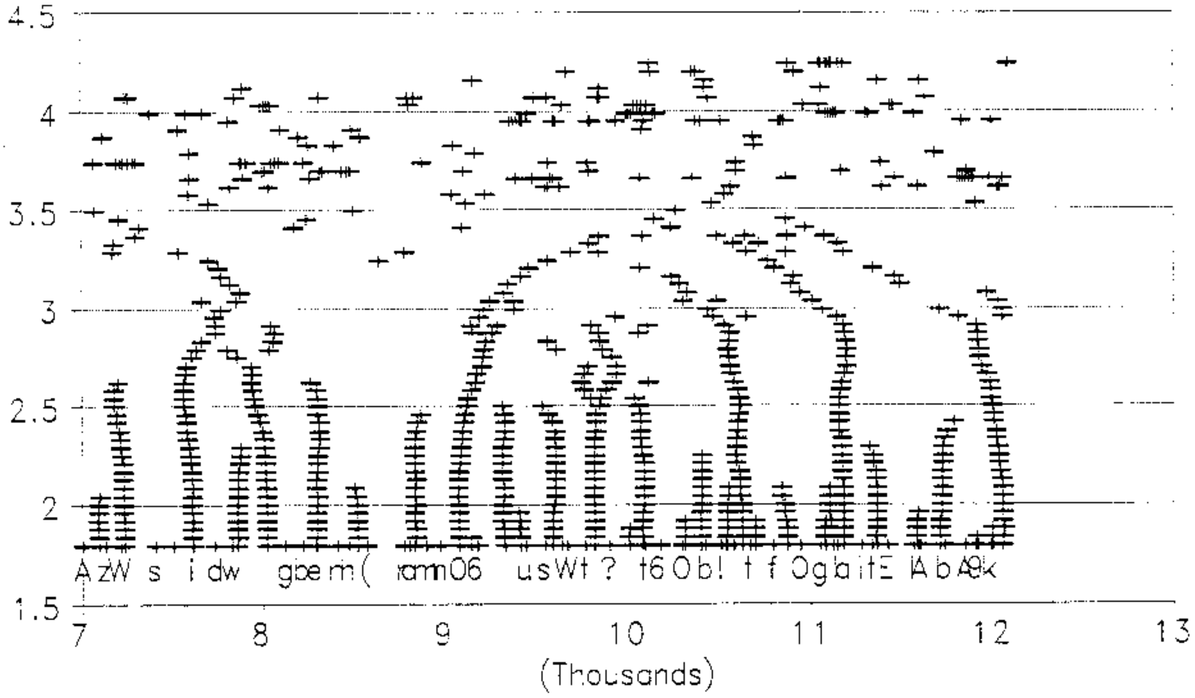


fig.15

This structure is perhaps even more unambiguous than the previous ones, but it contains one serious difficulty: The adverb is not clearly adjoined to the rest of the syntagm (although it is possible that a more dense analysis including more maximum points could have revealed this connection). Except for this, the adverb structure is very similar to the one in figure 12. The numeral 'három', which in the previous xz-diagrams and in all the above xy-curves had an outstandingly clear segmentation, is here split into two parts which do not even extend particularly high up. Evidently, the impact of stress on the continuous syntactic structure is rather important. This becomes even more clear in the verb phrase, which now seems just as undisputably <u>leftbranching</u> as it was rightbranching in the previous diagram. A binary branching abstraction of the object, the verb and the subject could look as follows:

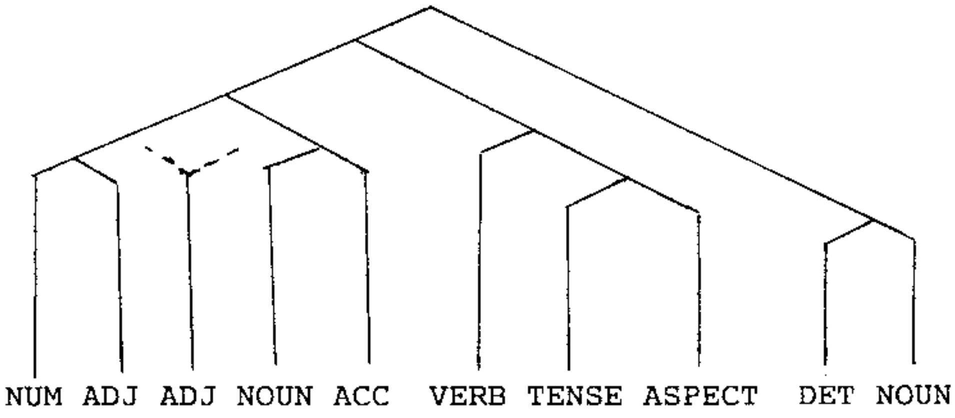


fig.16

When the first two sentences in the syntagm is defined in the same way, i.e., with a particular representation of the stressed vowels, their structure is as follows:

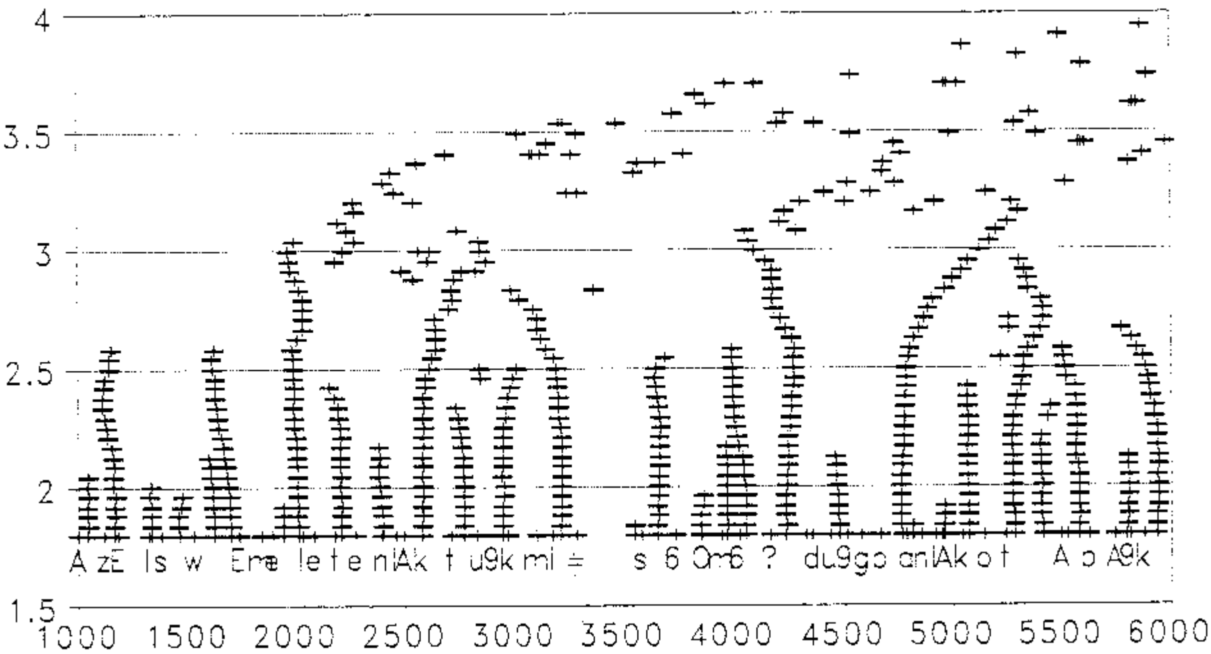


fig.17

In this representation, some of the words have got an improved structure. This is most notably the case with the noun in the adverb phrase. 'Emel-et-en' consists very clearly of three separate branches. In the above xy-curves as well as in the previous xz-structures, this form has for the most had only two peaks, with a 'double exponency' in the latter. The preterite '-t-' in the following verb is also much more clearly segmentable, and inclines leftwards towards the root. The following person marker '-unk' appears to be inclining rightwards towards the subject 'mi', although there is a couple of points on its lefthand side which could indicate a split and adjoinment to both the left- and the righthand constituent. As mentioned above, this would not seem ungrammatical: The personal pronoun subject 'mi' expresses exactly the same as the person marker '-unk', and the first serves normally to emphasize the second (although it is not properly optional in the present context). A structure where these two are joined into a single constituent must therefore be considered wellformed.

To the right of the middle pause sign, the most notable feature is that the INESSIVE '-ban' has not been improved: On the contrary, it incines more clearly than ever towards the righthand constituent. It may in this context have some relevance to note that this particular suffix is at the moment about to lose its distinctiveness relative to the suffix '-ba', with which it tends to be neutralized in current Budapest usage. This loss of the distinctive opposition by the disappearance of the final nasal would obviously have been notable in a corpus of speech sounds, and would have caused a somewhat different structure over the present syntagm.

The next diagram is generated by the distributional properties of corpus E, which means that it has identical conditions with the structure in fig.15 above, except for the additional introduction of phonological rules across all word boundaries:

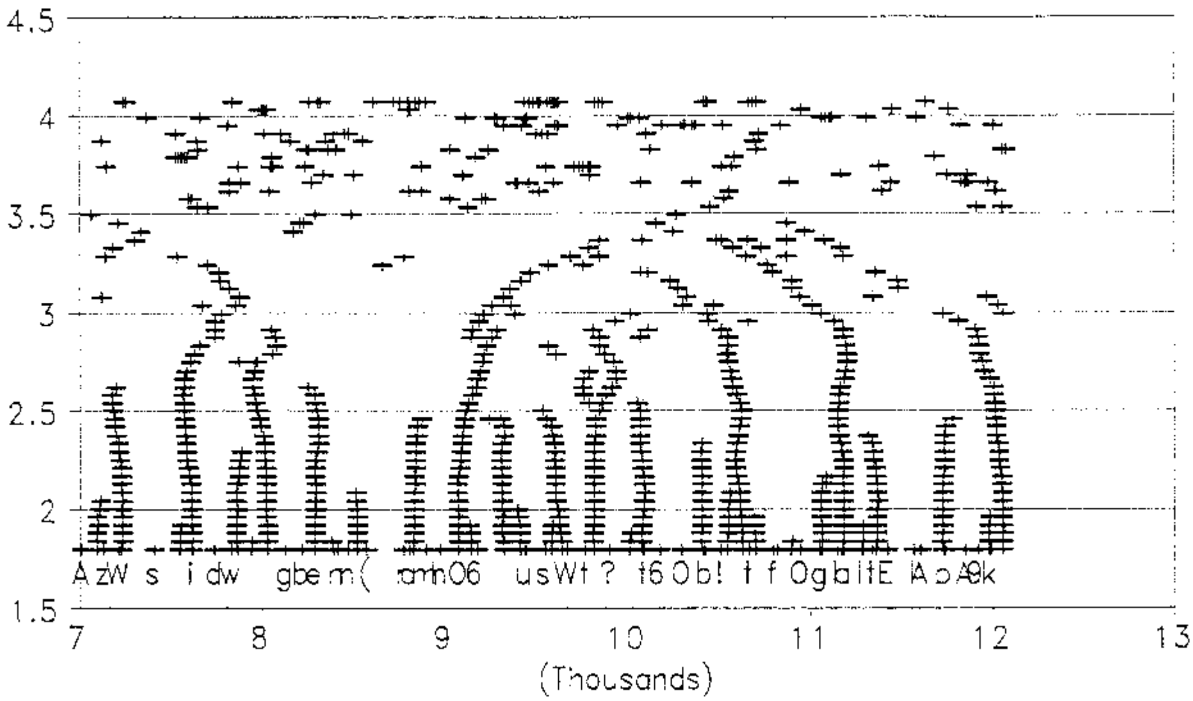


fig.18

As is seen from a comparison with fig.15, the differerence is very small indeed, and seems to be most notable in the lower parts of the diagram. If this is a general tendency, then we can assume that word-external phonological processes may have its most notable impact on the segmentation at lower levels (e.g. morpheme level), while the impact is not serious on the overall non-discrete syntactic structure.

In short, the definition of the corpus has a very notable influence on the continuous syntax. This seems to be the case in particular for the vertical resolution. It can be assumed that a radical change in definition may also give a radical change in parts of the syntactic analysis. We have, though, only scratched the surface of the definition. In particular, we do not know what impact a much better resolution may have on the structure, but it seems reasonable to assume that this may seriously improve the syntactic analysis in some cases, if the structures we have investigated is in fact a part of the linguistic competence of speakers.

Some important questions come immediately to mind:

1. Do the syntactic functinactually express grammatical structure? Will the second syntactic functionshow the same form?

2. Are there any additional factors - besides the possible grammatical ones - which could cause the diagrams to obtain the observed structure?

3. Is it possible that the structures could be secondary effects of the impact of grammar on the distribution of sounds? That is, could these structures be the traces of grammar rather than its representation?

To begin with the interpretation of the syntactic function, the function we have employed here suffers from the disadvantage of being not immediately interpretable. It is a fairly easily understandable representation of how for example morphological units relate, but it is less obvious that this also pertains to pure sound units with no meaning attached to them. To arrive at a more principled interpretation of the function, we will introduce the second syntactic function, such as it was described above, in its form D(a,b) = I(a) - I(a|b). As was mentioned above, this can be fairly easily related to the information theoretically classical definition of structural dependency. The following curve shows the information dependency structure of the first part of the syntagm under investigation:

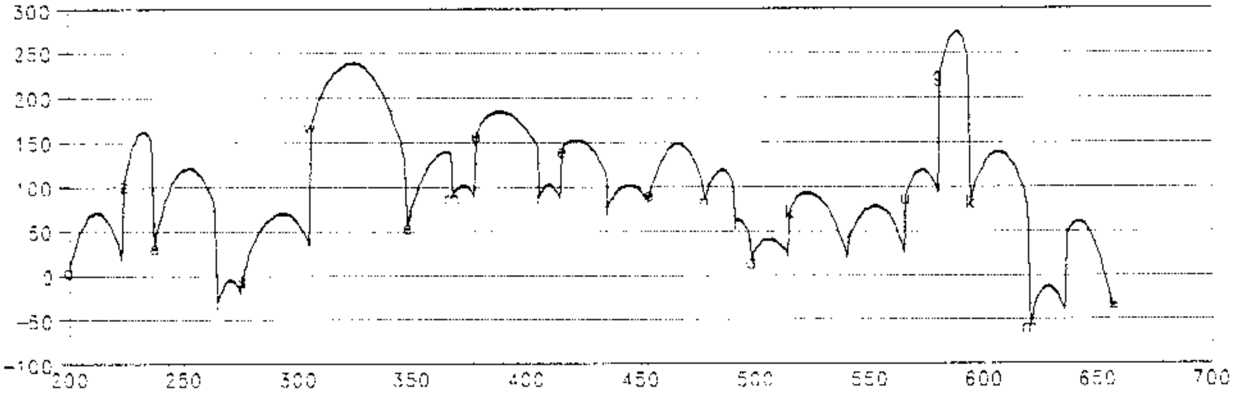


fig.19

The curve shows some notable properties of this function. First of all, it points to the obvious insufficiency in our corpus redefinition and the very rough vertical resolution. This becomes visible in the sharp breaks on the segment borders: Since the vertical resolution is poor and the horisontal relatively good, each segment will be represented by a considerable repetition of the same symbol. This gives a very low surprise (information) value at the middle of the segment, and a higher information value close to the border. The dependency, defined by the information value reduction caused by the conditional information value, will thus be high in the middle and low at the boundaries. These sharp break at the borders would of course not have beenso prominent if the vertical resolution had been much better and we could distinguish between a large number of different sound qualities. In this case, the surprise (information) values would have been smoothened out, and the curve would not have shown such notable breaks. But it would still basically have shown the same rising at the beginning of a segment and a falling towards the end.

Therefore, with a better vertical resolution, the curve will be smoother, but we will still find the local maxima approximately as they are in the present curve. Thus the phonemes can be defined as local maxima in the informationally defined dependency between very small sound segments.

The following diagram shows the first sentence in the syntagm, with the curves smoothened by averaging. Each measurement point has been averaged with the five points to its left and four to its right, such that the y-values in the following curve is the average over ten measurement points á 5 ms, i.e., the average informational dependency over 50 ms:

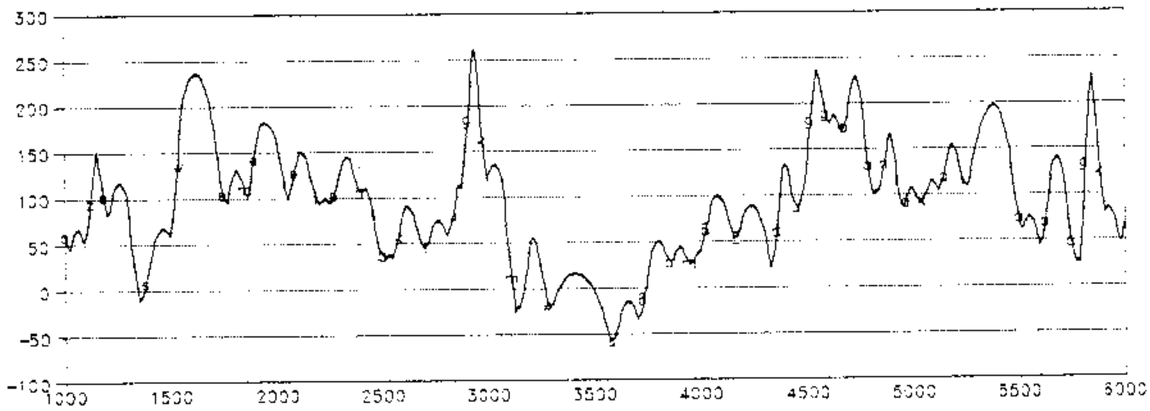


fig.20

The same fundamental relationships as in the pervious curve can be recognized. What is of interest now, is the degree of similarity with the curve over density fluctuations. The following curve is basically the same as in fig.6 above, only extended to cover the whole sentence:

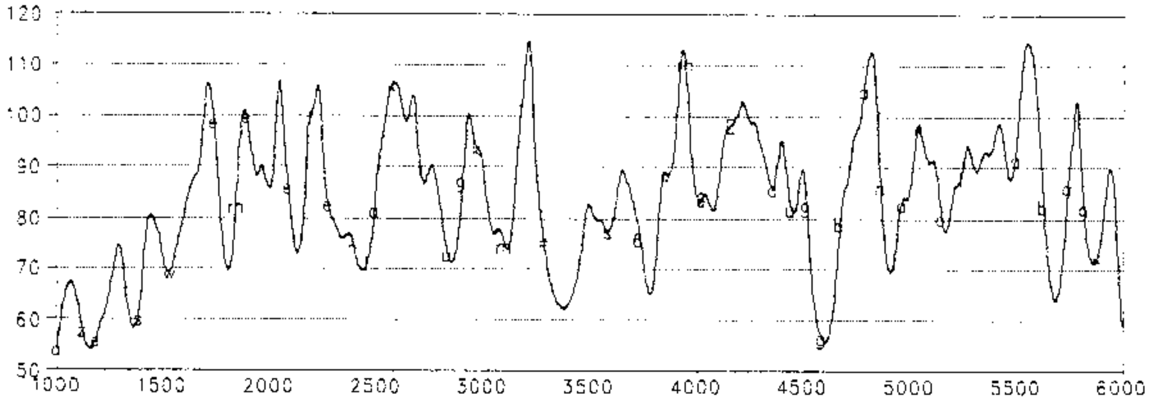


fig.21

The beginning of this curve is not fully reliable, up to approx. 1800. The curvatures in 'emeleten' do not correspond fully to the cuvatures in fig.20, although the end of the word looks similar: The 'n' is associated with a small bump. Next, the morpheme 'lak' starts more to the left in fig.20 than in fig.21, but except for this, both this morpheme and the following preterite 't' seem to be associated with the same kind of curvatures. This is also the case for the following possessive '-unk' as well as the pronoun 'mi' before pause. The difference is mainly a matter of positioning of the symbols on the line: If the symbols in fig.20 had been moved slightly to the right, the match would have been almost perfect. As mentioned above, this has to do with the problem of which part of the time line is associated with which symbol. For the density fluctuation curve, we chose to put density value on the middle of the interval over which density computation takes place. The same has been done for the averaged interval in fig.20, although it is less obvious that the solution is the right one in this case, and it may in general well be that there is more or less delay (psycholinguistically) in the assignment of y-value to the sound qualities on the time line. This problem is exactly what is exemplified here, and the correlation with the curve in fig.21 may suggest that the sound quality symbols in fig.20 should in fact be moved slightly to the right.

We also note that the heights of the curvatures are not the same in the two curves, but, as we will return to below, this seems not relevant for the present discussion. Not only are the sources for the investigation too full of errors to draw any information from the actual y-values, but we have also for the present concentrated on the local maxima and minima, and for these, it is not relevant whether the curvature is high or low.

The rest of the sentence can be compared, and similarities between the two curves can be detected to a smaller or larger degree. The similarity may be easier to trace if the averaged interval is increased, as in the following curve, where the averaged interval is 50 measurement points, i.e., 50 * 5 ms =

250 ms interval:

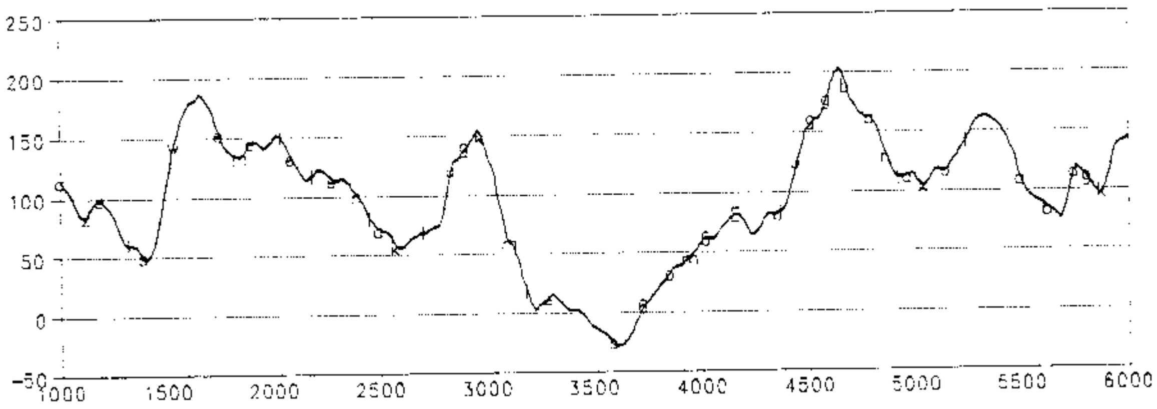


fig.22

This can be compared with the curve in fig.3 (interval: 70000). The main tendencies are clearly the same, which can also be concluded from a comparison of the same curves for the next sentence in the syntagm:

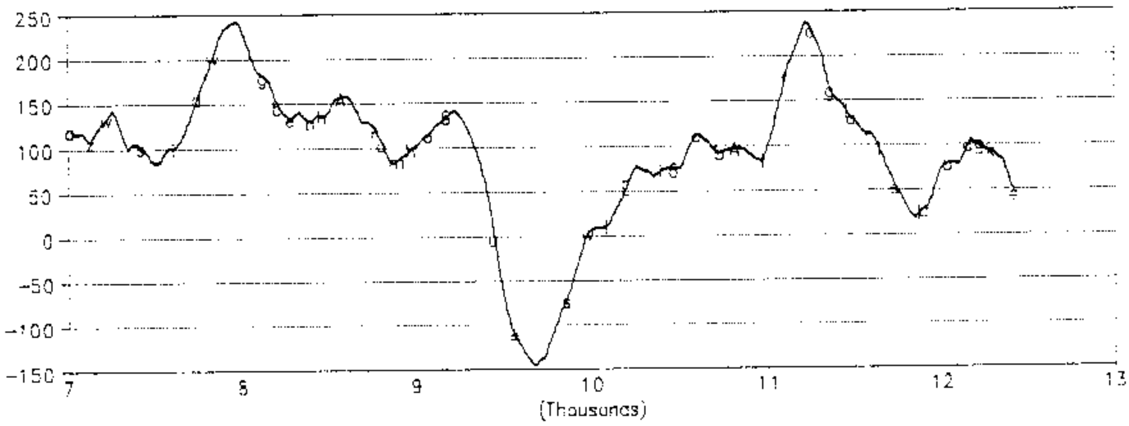


fig.23

This seems in general to be the case, and it points to the conclusion that the two functions describe very similar properties of the syntagm. For our purposes, the most interesting consequence of this is that even the density fluctuation curves must basically cover the grammatical structure of the sentence in the same sense as the informationally based curves, and this is within traditional information theory (see e.g. Krippendorff 1986) a well established measure on distributional dependency.

The question naturally arises whether we can abstract a branching tree from the informationally defined dependency curves in the same manner as in the above diagrams. The answer to this is most probably 'yes', but the framwork of the present study has not allowed for a more thorough investigation of these structures. In particular, it seems difficult to average the curves in a reliable manner, mainly due to the poor vertical resolution of the corpus definition. I have tried out a few possible approaches, and the method which seems to approximate a tree structure most closely under the present conditions is the following: 1. All dependency value variation is levelled out within the interval of identical segments (i.e., all local curvatures in fig.19 are transformed into horisontal straight lines). This is done to reduce the sensitivity of the function and the impact from the poor vertical resolution. 2. Going from left to right in the syntagms, it is counted how many such y-values must be added (cumulatively) to reach a certain limit L. The inverted number of such y-value additions is a measure on the relative amount of information concentrated in the area under investigation. 3. This inverted number is assigned to the middle of the interval of added y-values. When going from left to right and performing this computation on each point, the x-value as well as the limit L is recorded if the inverted number is (roughly) a local maximum. 4. The whole procedure is performed in loops with increased value on the limit L untill a final maximal value is reached. The following diagram shows the result of one of these tests. It is over the first sentence in the syntagm under investigation, and should be compared with the above diagrams in fig.8, 10 and 17:

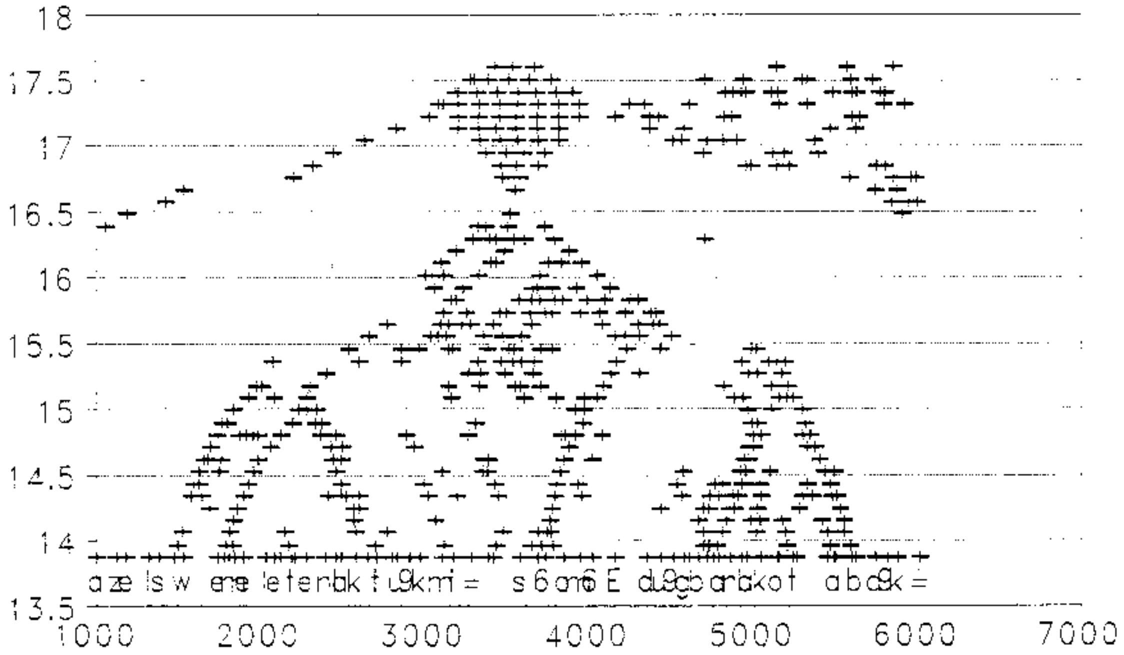


fig.24

Although there evidently is structure here, and it is likely to be somehow related to the grammar of the sentence, it is not as immediately recognizable as the structures generated with the first syntactic function above. There is, though, one main branch from each main word in the sentence. They are also basically correctly connected (as compared to the expected structures), and the same fundamental relationships between the braches as in the above curves 8 and 10 can be recognized. The next sentence in the syntagm is depicted in the following diagram:

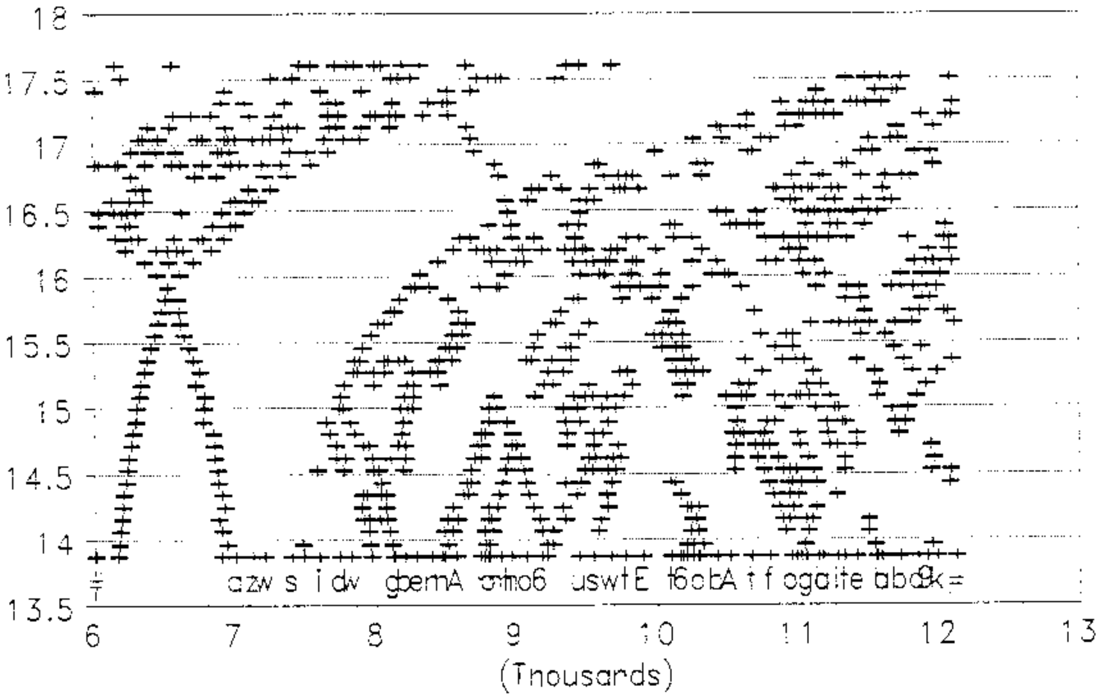


fig.25

When compared with the structure in fig.12, we can again see some of the same shapes, such as the righthand branching over x = 8 (the adverb), the branching between 'három' and hosszú' over x = 9, but the rest is more opaque. Although there evidently is a longer distance from this representation to traditional binary branching structure, it is still possible to recognize some of the basic grammatical structures.

This shows perhaps more than anything else the large importance of the syntactic function and its huge influence on the structures which can be generated by means of the distributional properties of speech sounds recorded in a database.

Secondly, it illustrates the importance of the function mapping from the dependency computation to the segmentation process and the extraction of the branching structures. It may well be that a more sophisticated computation of the xz-plane than have been presented here can yield a better approximation to a syntactic structure. (For the first syntactic function, this amounts to the density functions).

As to the second question we posed above: Can there be other reasons for these structures than purely grammatical ones? As we have seen, the corpus redefinition is of crucial importance for the resulting structure, and this is particularly the case for the treatment and representation of silence. The structure may be considerably changed if we add or remove a pause equalling 2-3 average phones of duration. This points more than anything to the large impact of the lack of variation in the corpus on the structures. In every position where the orthographical conventions prescribe a comma, the simulated speech will contain 300 ms silence. This gives a massive regularity to the representation of silence. We may now ask whether this phenomonon of over-regularity in the simulated speech can be the reason for at least parts of the structure? The answer is probably no. In the initial stages of the investigation, a very simple time metric was used: A short sound counted as one segment, a long sound as two. As the duration values were subsequently improved by the introduction of context-dependent duration of phones, *the segmentation was also improved*. The local minima coincided then better with linguistic boundaries than what they had done earlier. This can be interpreted in two ways: 1. The introduction of variation will in general improve the structure. (Of course, not any variation will do: It must be variation within the boundaries determined by speech performance). 2. When phones have their duration determined by the environment, this will increase the variability of the code, but it also *adds dependency* to the sequence of symbols. This is due to the nature of the imposed variation: It must necessarily be rule-governed variation (unless we design some random-generated variation within certain limits). This means that since, say, a sound [t] immediately in front of the sound [e] always has the duration d1, which it has in no other environments, this particular duration of [t] will make the [e] more predictable and hence distributionally bound to the [t]. Similarly, in the vertical dimension, the velar nasal appears only in front of velars, which make the latter very predictable from the former and vice versa (as can be seen in e.g. fig.19). Thus, when we introduce variation by rule, we impose grammatical structure on the code. Clearly, this is not irrelevant for the present question. It does, though, not mean that we have introduced non-existent structure on the code, although speech performance may show more or less deviations from the average values we have utilized in the symbol and time definitions. What it does mean, is that *phonological rules may have a core function in the distributional coding of the syntax*. Thus phonology may have a *function* in the overall grammatical system: it is not there only to impose a limitation on the inventory of perceived sounds, or the like, but may be the very code by which syntax is perceived and learnt.

For our purposes, it means that we have introduced one set of rules. Speech will evidently deviate from this, but its average realization will cluster around these values, and the increased predictability of sounds (environmentally conditioned) in the present code is therefore most probably nearer to the linguistic reality of speech performance than the initial code which had a simpler time definition, less variation, and thus less intersegmental dependency.

Therefore, to answer the question, it cannot be the lack of variation which causes the observed structures. If there is 'artificial' structure present, it must be due to the imposing of more rule-governed variation in the code than there is in actual speech. Whether this can be the case or not, cannot be determined right away, although it seems fair to guess that this hardly can be the case (for the very simple reason that such artificial structure would not conform to grammatical structure). As repeatedly mentioned, the final test on the reality and presence of the observed structures will have to be made on a corpus of actual recorded speech.

The third question above concerns the possibility that the structures can be secondary traces of syntax rather than a primary cue to it. We may well think of a grammar which works entirely independently of distributional properties, i.e., with no regards to the regularities in the expression side, but which happens to create a phonetic surface which possesses distributional dependencies as a bi-product. This would for example be the case if the syntax contained strong constraints on word class order, and word class was (additionally) signified by overt morphological markers or even by morpheme structure conditions. (Note that this is normally not the case: There is typically an inverse proportionality between constraints on word class order and morphological marking). If such were the case, then we could possibly read some syntactic structure from the distribution of sounds. This can, though, hardly be the case for Hungarian. As is well known, constituent order is extremely free in Hungarian, and, as shown above, even the morphological markers are extremely homophonous across word class boundaries. These two factors are probably sufficient to exclude the possibility for the present structures to be secondary arbitrary traces of a syntax which works independently of sound. It does not imply that the syntax works by means of sound distribution, but it probably means that sound is relevant to the Hungarian syntax.

Thus there seems to be good reasons to assume that the observed structures are indeed a part of the grammatical system of Hungarian. This is, of course, on the condition that the distributionally conditioned structures do in fact conform to the grammatical structures. The syntagms covered by the present investigation (not reported here) have shown a relatively large degree of correlation in structure, and only very few major cases of serious segmentation errors (i.e., at a lower level).

For a more extensive account including a larger number of syntagms, we will refer to the analyses presented in the appendix A below.

To the reading of these, we will finish this chapter by reviewing for the reader the possible error sources which may be part of an explanantion to the structures:

1. The composition of the corpus. Although the syntactically relevant data are extracted from very local sound relationships, the composition of the corpus is not entirely irrelevant for the present purpose. The corpus which has been

used does not contain many of the characteristic contractions of oral language, nor the characteristic vocabularies, in particular those in the dialects of motherese.

2. The general lack of variation, at all linguistic levels.

3. The very poor vertical resolution, which hardly exceeds the size of a typical phoneme inventory. Very important.

4. The insufficient data on durational values, and in particular the incomplete definitions of environment. Only the righthand environment is included, and even this is very limited.

5. The errors in the corpus text (abbreviations, foreign names and phrases, typographical codes etc.).

6. The choice of syntactic function to measure dependencies. Very important.

7. The choice of density function (for the present syntactic function).

8. The choice of x-value to which a dependency value is assigned. In the present investigation, a density value is assigned to the symbol with the same x-value as the middle of the interval for which the density has been computed.

9. The choice of the single symbol model rather than the multi-layered model.

10. Finally, and most importantly, the diagrams contain only a very small part of the information actually contained in the syntactic structures. They show only the presence of local y-value maxima, and does not contain anything on neither the actual y-values nor the relation between the these in e.g. adjacent peaks.
zz

CHAPTER 3: THE COGNITIVE BASIS FOR NON-DISCRETE GRAMMAR

Zellig Harris, whose approach to linguistic theory is very akin to what we have presented here, writes: "Does [the distributional] structure really exist in the language? The answer is yes, as much as any scientific structure really exist in the data which it describes. [...] Does the structure really exist in the speakers? Here we are faced with a question of fact, which is not directly or fully investigated in the process of determining the distributional structure. Clearly, certain behaviors of the speakers indicate perception along the lines of the distributional structure" (Harris 1955 p.149). What we have found in the previous chapter seems to suggest an affirmative answer even to the second question, although the distribution which Harris writes about is of a kind somewhat different from what we have investigated. Since the work of Harris, probabilistic grammar has been very much overshadowed by the prevailing generative paradigm. This is not the least due to the early rejection by Chomsky of any possibility for a probabilistic grammar to describe the grammatical architecture of linguistic competence. (See e.g. Chomsky 1956 and 1957). Most discussions have, though, been centered around the possibility of a *Markov grammar* based on the conditional probabilities of *pre-established linguistic units*, in particular a grammar over the probability matrices of word occurrences. This is in essence a continuation of the work of Harris, and the distibutional structure which he talks about, is a structure based on the distribution of discrete linguistic units. The Markov grammar models which have been presented are all founded on discreteness of the units. The failure to set up a satisfying model may be closely connected to this precondition.

What is basically new in the present approach is the continuity of the input to the grammar, which implies that the phonetic qualities of sound frequency distribution and intensity as well as the time metric are the significant parameters to the syntax. It makes non-discrete probabilistic syntax an essentially *perceptual process*, and the syntactic structure is the output of a perceptual organization of incoming sensory data. This need not imply that this perceptual structuring is the only syntactic structure in competence: There is nothing in the model which prevents a subsequent processing of the structures in e.g. a *generalization process* or in the application of a possible *transformational component*. Rather, the model contains an account on how a primary syntactic structure can be extracted from the sensory data, to function as the input for a possibly more logically oriented grammatical competence.

For our purposes, what is of interest is whether the structures we have found in the data can be directly perceived. If this be the case, then we can assume the distributionally based syntactic structure to be directly accessible to language users. The core function in this process will have to be the gradual accumulation, through exposition to linguistic data, of an expectation of sound co-occurrences. This is the psychological pendant to conditional probabilities: A mathematical expression such as p(a|b), the probability of the symbol 'a' to appear under the precondition that the symbol 'b' has occurred, is a measure on such expectancies. The internal representation of a conditional probability is in the form of a learned capacity to predict the next event in a series of interdependent events. If a subject is asked to guess the next event in a series of trials, and each time the system is in state A of a Markov chain the subject's guess for the next state is recorded, and it is found that the distribution of the subject's guesses is in accordance with the transition probability distribution in state A, then we will say that the subject's expectancy is a representation of the true transition probabilities. It has been repeatedly shown experimentally that learning is a process in which the expectancy distribution of a subject approaches asymptotically the transition probability distribution in a Markov chain. All current theories within mathematical psychology on learning processes are stochastic and emphasize the probabilistic nature of learning. Coombs et al.(1970) refers to the two main models as the operator model (initiated by Robert Bush and Frederick Mosteller in the beginning of the 1950's) and the finite state model (William Estes 1950) respectively, and both are basically on a Markov form: "The two developments formulate the probabilistic nature of the learning process in the same way. The process is conceived of as a sequence of discrete trials. Each trial consists of the presentation of a stimulus situation to which the subject responds by selecting one from a set of alternative responses in accordance with an associated set of probabilities; the response is followed by an outcome, which may induce changes in the probability values before the next trial. Therefore, in brief, the learning process is analyzed into a sequence of discrete trials, each of which consists of a stimulus, response, outcome, and resultant probability flow. All models are concerned with describing this flow of probability from trial to trial and the resulting sequence of distributions" (p.259). Thus, it is fully in line with learning theories to assume that, by a learning process, the psychological expectancy of co-occurrence may be seen as approaching the real, measurable transition probabilities in the acoustic data which the subject is exposed to. One part of the linguistic competence which is built in the initial stages of a language acquisition process is therefore constituted by a knowledge of the conditional probabilities of sound co-occurrences (as well as the much simpler unconditional probability values, which amounts to the plain frequencies of sounds). These data corresponds to the database in our investigation.

For the perceptual process, a question of central importance is how this knowledge is employed in the perceptual organization of the incoming sensory data. There are theories which directly employ the information theoretical apparatus. Garner (1962) is the classic on an information theoretical approach to perceptual processes and discrimination. It provides a model for, amongst other things, pattern recognition and concept formation. It is, for practical purposes, mainly concentrated on visual perception, but do also cover auditory perception which may be treated in very similar way. The notion of uncertainty, measured as entropy, is central to the model, and it analyzes perceptual organization as a function of information flow: The higher the uncertainty (i.e., the entropy as an average information value) in some area of the sensory field, the more prominently will it appear in the perceptual process. Numerous experiments are reported, among which the following may be relevant in the present context: "The beginning and ends of words carry the greatest information, and the middle letters of words are the most redundant. Does this fact have any effect on how words are perceived? Data from an experiment by Haslerud and Clark (1957) show that it does. They required subjects to read nine-letter words which were presented tachistoscopically for 40 msecs at a rather low level of illumination. The accuracy with which various letters were reported is shown [in the following figure:]" (Garner p.259f.).

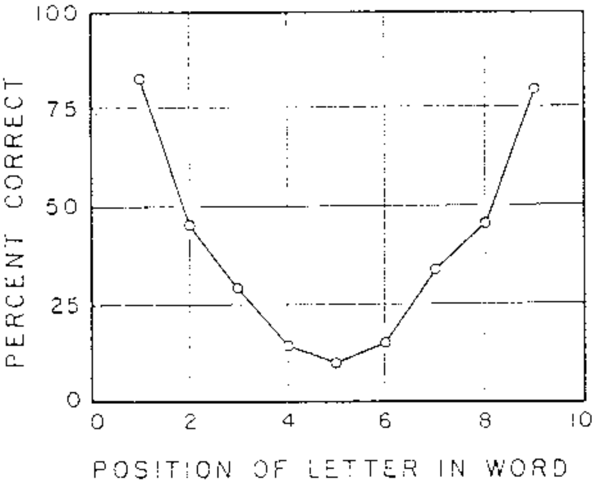


fig.26

(The figure is after Garner p.260). What this experiment shows, is that the the beginning and end of (nine-letter) words receive the highest perceptual attention. As was found in the curves based on the informationally defined dependency function, the dependency, i.e., the constraint on distribution, was generally highest in the middle of linguistic units in the Hungarian data, which means that the uncertainty or entropy, i.e., the information flow, is generally highest in the beginning and the end. The experiment reported in Garner shows that this (distributionally defined) information flow is directly perceptible in the sense of governing the level of perceptual attention. This experiment illustrates the basic content of informationally oriented models of perception. Fred Attneave, who has also contributed substantially to the information theoretically oriented work in the psychology of perception, presents e.g. in Attneave (1954) an informational interpretation of visual perception: "Information is concentrated along contours [of sensed objects] (i.e., regions where colors changes abruptly), and is further concentrated at those points on a contour at which its direction changes most rapidly (i.e., at angles or peaks of curvature)" (p.184). The point is that these areas of high information flow will also be crucial for perceptual discrimination and pattern recognition, a proposal which seems to be in line with other models of perception.

Garner reports another experiment on the effects of distributional constraint on the perception: Miller (1958) generated nonsense words of length 4 to 7 letters by means of a restricted alphabet (only four discrete symbols: the consonants 'g', 'n', 's' and 'x'), and subjects were asked to learn and subsequently recall lists of nine of these nonsense words. There were two types of words: In one set, they were generated by means of distributional constraints (i.e., corresponding to morpheme structure conditions) which overrepresented some letter sequences and underrepresented others, and in the other set, the words were generated with no such constraints, i.e., they had a random distribution. The result is as shown in the following figure (after Garner, p.272):

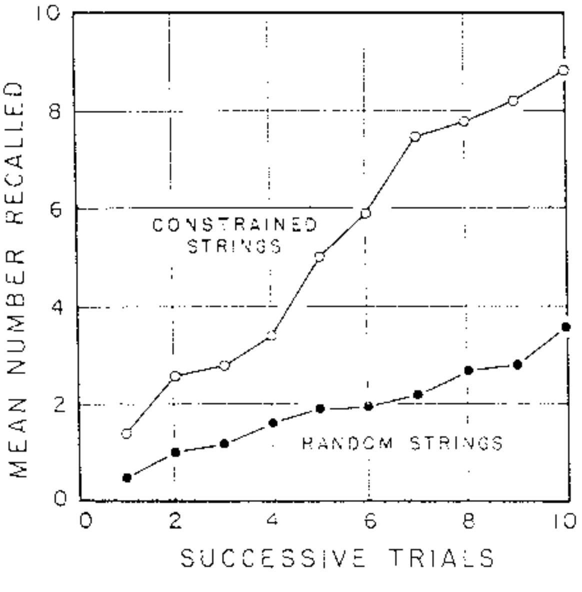


fig.27

The y-axis shows the mean number of correctly recalled words, while the x-axis shows the number of repetitions, i.e., roughly, the rehearsal time or amount of learning. The experiment shows that distributional constraints highly favours the learning process. In the present experiment, the alphabet size is so small that co-occurrence constraints are extremely rapidly learnt, and a database of transition probabilities will be very quickly established.

However, a random distribution has a higher information flow than a constrained one (distributional constraints are information flow filters). This means that there is more information in the random strings than in the constrained ones, and if the results are corrected for this (the exact information flow can easily be calculated), it turns out that more information is acquired per time unit through the learning of the random strings than through constrained strings. "While it takes longer to learn random sequences, the amount of information gained per time unit is greater than for constrained sequences" (Garner p.273). Thus, again, the information flow affects the perceptual process in the sense that the perceptual organization operates directly on the structural information. If experience is acquired, such that an expectancy matches the occurrence of symbols, this amounts to a reduction of the information flow, while the learning rate is improved.

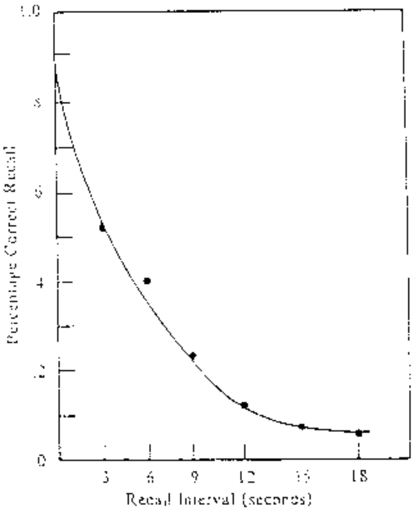
This tells us two things: The perceptual organization of incoming sensory data, i.e., the perceptual grammar, is concentrated on peaks of information flow, i.e., where there is minimal distributional constraints, such as we typically have found around segment borders. The learning process is, though, favoured by maximal distributional constraints, such as we typically have found in the data between segment borders. Grammatical processing by rule should thus be favoured by low distributional constraints, i.e., high perceptual salience due to high information flow, while lexical representation by rote learning should be favoured by a lower perceptual salience and a lower information flow through constrained sound sequences. The (according to my knowledge) universal presence of morpheme structure conditions fully supports this.

This suggests that stretches of sound over which there are perceptible distributional constraints will tend to be lexicalized, while the absence of such constraints suggests an interpretation according to the grammatical structure of the sound sequence. As the experiments reported by Garner point to, the structure will be established on a perceptual basis.

It should finally be mentioned that the informationally oriented models of Garner has been somewhat refined and updated in Garner (1974).

We do, though, not have to rely on information theoretical measures on distributional structure. Although the first syntactic function in our investigation showed notable similarities with the informationally defined dependency curve, (and may thus indirectly receive support from the Garner model), the functions are still basically different. The main difference between them lies in the fact that the time distance between the pairs to a larger extent is an integrated part of the first syntactic function (and not only of the records in the database). The most immediate reaction to this function, from a perceptual processing point of view, is that it involves large amounts of data. The values computed for the y-axis, continually along the x-axis, are based on density computations over hundreds of thousands, or even millions, of points, such as we have defined the function. This is, though, not necessarily large amounts of data in a perceptual framework: The amount of information which reaches the sensory register from the senses is so vast, that one of the main problems for perceptual psychology is to account for how it is possible that such immense amounts of data can be organized and structured into a conceptual framework with such unbelievable efficiency as it has in the human perceptual processing.

A particularly interesting notion in this connection is the by now generally approved distinction between a short-term and a long-term memory system, in its modern form experimentally tested simulateously by Brown (1958) and Peterson and Peterson (1959), although it in its original form stems as far back as to William James. Experiments showed that when subjects were presented with a small set of nonsense words, and they subsequently were asked to perform rehearsal prevention tasks, such as counting backwards, i.e., they were prevented from rehearsing for themselves the data which they had been presented to, and they finally after a relatively short interval (from zero to 18 seconds) were asked to repeat the nonsense words, it turned out that the data decayed extremely rapidly from memory. The following figure shows the amount of data retained in memory as a function of time:



(The diagram is from Ellis & Hunt (1989) p.70). Since such rapid decay evidently is not the case for all memorized data, it was concluded that there must be several (at least two) different memory systems with different functions and coding systems. The main hypothesis was that short-term memory had its basic function in keeping relatively large amounts of data in a memory store for perceptual structuring before the readily structured 'chunks' of information are handed over to the long-term memory, which has a virtually unlimited storing capacity. For this purpose of functioning as a central processing unit, there is no need for a long storage capacity in short-term memory, since the data will continuously be replaced by new incoming data.

Subsequent experiments have corroborated these findings. In particular, Baddeley (1966b) reports some experiments in which subjects were presented auditorily with lists of words (not nonsense, and short enough to be retained in short-term memory) which were either acoustically or semantically similar, and the subjects were asked to recall the words after a certain retention interval. The remarkable results of these experiments showed that the lists containing acoustically similar words were poorly recalled (only 9.6% correctly reproduced, against as much as 82.1% for control lists), while the lists of semantically similar words were reproduced with 64.7% correctness (as against 71% for control lists). From this and similar experiments testing the long-term memory system, Baddeley concluded that "subjects show remarkable consistency and uniformity in using an almost exclusively acoustic coding system for the short-term remembering of disconnected words. There is abundant evidence that this is not true of long-term memory" (Baddeley 1966b). On the contrary, Baddeley found that long-term memory had a semantically oriented coding system.

Thus, in addition to the basic sensory register which receives the raw sensory data for a first preprocessing, the model contained one memory system (of short duration) with an acoustic coding and another system with a semantic coding (and a long duration).

These findings are by now generally accepted, although the sharp distinction between a short-term and a long-term memory system has been somewhat softened. Baddeley later revised the notion of a particular short-term memory system with a more general notion of a working memory with a number of different functions, and Baddeley (1986) presents a model of the working memory as consisting of

1. a modality-free central executive (something close to traditional attention),
2. a module specialized for spatial and/or visual coding,
3. an articulatory loop which holds information in a speech-based form.

The latter consists of a passive phonological store which is directly concerned with speech perception, as well as an articulatory process that is linked to speech production. As to the former of these, speech-based information can be entered into the phonological store

1. directly through auditory presentation
2. indirectly through subvocal articulation
3. indirectly via phonological information stored in long-term memory.

(The previous passages are rough quotations from Eysenck & Keane (1990)). There is by now very strong evidence that the closer we come to the sensory register, the more will the coding be in the form of acoustic parameters (frequency, intensity, duration), while the long-term memory store contains mainly semantically coded information as well as the accumulated knowledge of sound co-occurrences.

The short-term nature of knowledge on acoustic data has been repeatedly shown by various experiments. Eysenck & Keane (1990) report experiments done by Treisman (1964), "who asked people to repeat back aloud (i.e., shadow) the message presented to one ear while ignoring a concurrent message presented to the other ear. She presented the same message to bnoth ears, but in such a way that the shadowed message either preceded or followed the non-shadowed message. When the non-shadowed message preceded the shadowed message, the two meassages were only recognized as being the same when they were within 2 sec. of each other. This suggests that the temporal duration of unattended auditory information in echoic storage is approximately 2 sec., although other estimates are slightly longer" (p.138). Furthermore, Baddeley, Thomson and Buchanan (1975) discovered that "their subjects could provide immediate serial recall of approximately as many words as they could read out aloud in two seconds. This suggested that the capacity of the articulatory loop is determined by temporal duration in the same way as a tape loop" (Eysenck & Keane p.143). Garner (1962) does not mention any particular time limit, but notes that "the most surprising aspect of laboratory experiments on the learning of statistical dependencies [in speech] is that what can be learned is so very limited. The available evidence suggests that simple contingencies of adjacent symbols can be learned fairly easily, but that longer sequences, even when invariable in nature, are learned only with difficulty. [...] Humans can learn not only distributional probabilities of a stimulus series; they can also learn sequential probabilities if these exist. The experimental evidence available suggests that learning of sequential constraints which exist over more than a very small number of steps is very difficult" (Garner p.305 and 308).

These findings are very much in line with our data. From the database extracted from corpus A, the amount of average distributional dependency per position (which for this database means steps of 5msec per position) can be computed as the sum of differences between the conditional and the unconditional probabilities. The following figure shows the relative dependency, i.e., the relative syntactic importance, such as we have defined it, in this database. The y-axis is the value of the function $y = \log \Sigma |p(a|b)-p(a)|$, while the x-axis is the time interval between the pair of sound qualities:

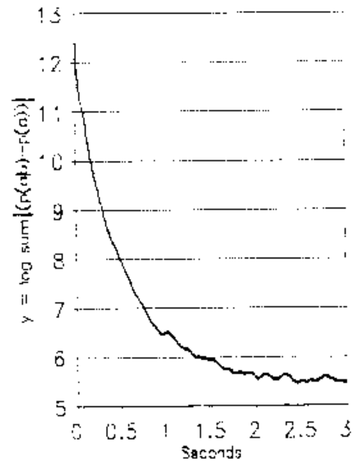


fig.29

What is particularly interesting in this context is the notable similarity with the amount of memory decay in short-term memory such as it is shown in fig.28, as well as the notable correlation with the 2 sec interval for storing capacity of acoustic data in the articulatory loop of the working memory model. What this suggests, is that distributional interdependencies between the sounds are indeed processed in the short-term memory system as a part of the perceptual processing of incoming acoustic data. This may be the very reason why these interdependencies are there, and in particular why they have the durational extension as they have.

The function of short-term memory is, as mentioned, to provide a readily processed 'chunk' of information which can be handed over to long-term memory for permanent storage in semantic form. If, though, this processing is dependent on the distributional properties (within approx. a 2 sec interval) of the speech sounds, it is evident that short-term memory (or whatever equivalent there is for it in a perceptual model) need access to the database over sound interdependencies. This is exactly what the revised model in Baddeley (1986) suggests, in which the articulatory loop has access to the phonological (speech-based) information stored in long-term memory. The problem we envisage here is the same as has been the locus of much controversy within perceptual psychology: It concerns the problem of whether perception is data-driven (bottom-up models) or concept-driven (top-down models), i.e., if the output of a perceptual process is ultimately determined by the incoming data only, or whether it relies on previous concept-formation and thus is influenced by long-term memory. For our purposes, the most important fact is that both opinions exist. The debate is by no means settled, but, as Eysenck & Keane (1990, p.95) remark:, "The most reasonable position is that the relative importance of bottom-up and top-down processing depends on the particular circumstances in which perception occurs. Most of the time, both kinds of processing will jointly determine perceptual experience and performance. As a consequence, what is needed is more of a theoretical understanding of the ways in which bottom-up and top-down processing interact". There seems to be much theoretical support for the assumption that short-term (or working) memory can have access to the knowledge of sound co-occurrence probabilities stored in long-term memory in the very perceptual processing of the acoustic data. This is also very much in line with the parallel distributed processing (P.D.P.) model, such as presented in e.g. Rumelhart and McClelland (1982), in which there is a feedback system which involves both incoming acoustic data and previously established conceptualizations. When talking about a perceptual process, they remark that "this [...] process was guided both by the acoustic features of the input itself and by top-down activation from higher levels through the word level to the phonological level" (p.92).

If we can assume that the knowledge of local sound interdependencies do in fact enter into the perceptual process, we may consider the hypothesis of perceptual parsing of the syntax as partly corroborated. The process will typically be a continuous assignment of syntactic dependency over a span of approximately (or maximally) 2 seconds, and a continuous parsing through a density (or any other appropriate) function. In the above diagrams, it was found that there was a prominent and clearly appearing structure up to approximately y = 3.5/3.6, which amounts to a time interval of approx. 3 to 4 seconds. Since this exceeds the interval which (according to Treisman's and Baddeley's findings) can be kept in short-term memory, we can either assume that the actual time span in which acoustic data can be processed is slightly longer than the suggested 2 seconds (such as hinted to by Eysenck & Keane), or we can assume that only structures extracted from intervals up to 2 seconds are syntactically significant. This amounts to reading the diagrams up to y = 3.3 only. The poverty of our data as well as the relative uncertainty connected with the definition of syntactic function do, though, not allow for any further generalizations or conclusions on this matter.

A final question concerning the amount of data to be processed needs perhaps to be discussed. How can a density function be handled in a perceptual processing? The situation for the syntagm parsing function is in fact very similar to a parallel process in the interpretation of visual data. Consider the following illustration:

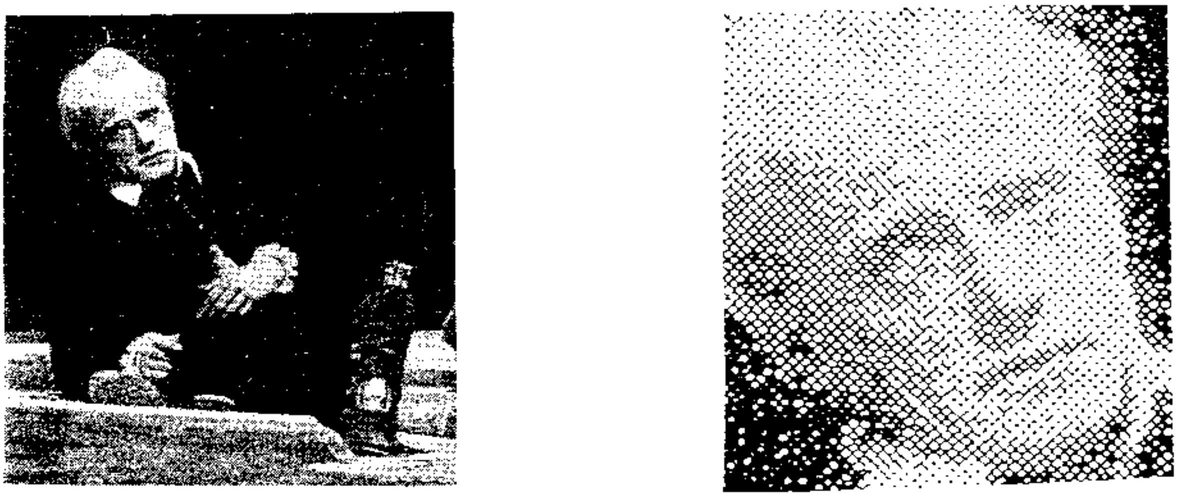


fig.30

(The picture is from Le Monde, 30.12.91, photo: B.Enguerand). When we see an ordinary newspaper picture such as the present, we have no problems in organizing the data into the recognition of a face, although the incoming sensory data are nothing but larger or smaller density of printer's ink on a flat surface. If the present picture should be digitalized and stored in a satisfactory form (with a sufficiently high resolution) in a computer, the amounts of data would have been very large indeed. For our purposes, the most interesting point is the way we interpret these density fluctuations. Corresponding to the most local density measurements in our investigaton, say, a morphemic level, we can also distinguish a local fluctuation in the picture's density as an eye, a lip, a wrinkle. When we parse it over a larger interval, say, the phrase level in the above curves, the local density fluctuations will merge into larger patterns, and we can recognize a face, a hand, a body against the background, and so forth. It may perhaps remain as a mystery how we can process such vast amounts of sensory data with such ease, but the fact is that we can. The acoustic parsing of speech may, such as the above syntactic function suggests, be in a similar form.

A final word on the apparently temporal nature of these density computations: This is, though, only apparently. Recall that it seems as if at least 2 seconds of acoustic data in succession can be held simultaneously in working memory. This means that for an interval of 2 seconds, the data will exist synchronically in memory, and can be processed just as similar amounts of data from any other sensory input, such as visual sensation. The continuously incoming data will update the memory as previously entered data decay 'in the other end', much as the landscape passes outside a train window. Thus the very capacity of the part of memory (presumably working memory) which is utilized for acoustic data processing will constrain the area over which dependencies can be recorded. If this capacity has an upper limit of 2 seconds (or slightly more), this corresponds well to the limits we have found for significant dependencies in the database as well as the limits for significant structure in the diagrams over local maxima in the density function. If this be the case, then we can possibly give a principled account for why phrases tend to have the extension they have, and why - although transformational grammar argues that sentences can be infinitely long without losing their grammaticality - phrases can hardly exceed certain limits of duration without losing their immediate comprehensibility. (Although, of course, much longer phrases can be handled and interpreted analytically by being kept in long-term memory).

Thus there is considerable support from perceptual psychology for the assumption that distributional structure is perceptible, and there are strong indications that it is processed by means of a short-term or working memory system, characterized by non-semantic, acoustic coding, such that the purely acoustic dependencies which exist in time can be kept synchronically for processing over a limited interval. In particular, it seems reasonable to assume that continuous grammar provides long-term memory with readily segmented chunks of data to constitute a lexicon of linguistic segments, to be semantically coded. This is fully in line with the generally approved conception of the encoding system in long-term memory: In addition to its particular susceptibility to semantic coding, a notable feature of long-term memory is that it seems to be of a distinctive or dicrete nature. Eysenck (1979) has found that "memory traces which are distinctive or unique in some way will be more readily retrieved than memory traces which closely resemble a number of other memory traces" (Eysenck & Keane (1990) p.151). This characteristic favouring of discreteness makes it similarly natural to assume that what we in the present study have termed discrete grammar operates on long-term memory, since it typically takes discrete items as input.

Discrete grammar generalized from continuous grammar.

In short, the continuous grammar which we have outlined here may have the general function of performing a basic, formal recognition task (of sound qualities) and providing the long-term memory with discrete segments for the lexicon. The continuous grammar will also generate a syntactic structure over the utterance. If a discrete grammar can be generalized from the continuous syntactic structures, then we have a possible principled explanantion for how a linguistic competence can be acquired through exposition to a language, and we have an alternative to the innateness hypothesis.

How can a generalized grammar be extracted from a continuous grammar? For the latter, it is essential that it will create a unique structure over all utterances, since it takes the acoustic parameters frequency, intensity and duration (with some fine resolution) as input, and we from a practical point of view can consider all utterances to be unique in this respect (it is hardly possible to create two acoustically identical utterances). This means that the segments which are the output from the perceptual parsing will never be exactly identical to any other previously parsed segment. If we consider the segmentation as consisting in recognizing local minima in the cognitive (mental) pendants to the curves we have presented above, then neither the curvature itself nor the sound sequence associated with it will be identical to previously segmented matter. In a generalization process, in which the first step will be to build a lexicon of discrete segments, the process of *cognitive organization* will be central. It is defined in Ellis & Hunt as "Organization is the process which groups discrete, individual items into larger units based on a specific relationship among the items" (p.92). In the present context, the discrete units will be the different sound sequences delimited by the local curvature minima. Organization typically favours *similarities* between the items which are grouped together, and is known to be of central importance for the functioning of long-term memory. But, as was mentioned above, long-term memory also favours *discreteness*: A distinct trace (i.e., which is maximally easy to distinguish from other traces) is more easily retrieved from long-term memory than a less distinct trace. This seems at first glance to represent to opposing conceptions of long-term memory. According to Ellis & Hunt: "We now seem to be confronted with diametrically opposed prescriptions for good memory. Organization argues for the encoding of similarities, whereas levels of processing [i.e., which pertains to the distinctiveness hypothesis] emphasizes encoding of differences" (p.104). This is, though, only apparently so: *Due to the very continuous nature of the matter to be stored*, in order for the 'chunks' to be sufficiently discrete or different, a process which organizes similar items into larger groups is of course necessary. An illustration of the process could be the representation of the set of real numbers as the set of integer numbers: All numbers between 1.500 and 2.499 are more similar to the number 2 than to any other integer, and if all these organized by this similarity into the integer number 2, (and the same is done with all other real numbers), the resulting representation (1,2...) will be more discrete than if they were stored as real numbers: There is more difference between 1 and 2 than there is between 1.499 and 1.500. (Evidently, over a certain interval the number of encoded integers will also be mich smaller than the number of real numbers). Thus the two long-term memory principles favouring organization by similarity and encoding by distinctness work in the same direction. A lexicon of discrete units may therefore be motivated by the very biological architecture of long-term memory.

The generalization process will, then, typically consist in the encoding of segments which have sufficiently similar curvatures *and* acoustic properties as identical lexical entries. There must be room for differences both vertically and horisontally. The limits for acceptable variation need not concern us here: It seems reasonable to assume that these will be conditioned by the particular (possibly subjective, see Ellis & Hunt p.93f) needs for distinctiveness, and may show considerable variation.

The syntactic structure generated in short-term memory may, though, be multi-dimensional, and segmentation can be carried out in a number of ways. In the function we described above, the structures were curving surfaces in three-dimensional space, and the choice of z-value would ultimately determine the segmentation of the utterance: A low z-value would give segmentations at morpheme (or even phoneme) level, while a higher z-value would yield word or phrase level segmentation. If, then, we assume that even these higher-level segmentations are encoded in long-term memory by the same principles of organization and distinctiveness, the lexicon will contain segments at all levels: Morphemes, words, phrases, and possibly phonemes. The discovery of the decompositionality of larger entries into smaller entries amounts to the discovery of a syntactic rule. This rule will typically be a phrase structure (or a morhological) rule.

We need not be concerned here by the way this rule will be represented cognitively, nor by the exact point at which it will become productive. In its simplest form, it need not be anything more than the discovery of the decompositionality, and the rule may be represented as such at an unproductive level. We must here emphasize that this decompositionality in its simplest form is nothing else than the recognition of the appearance of local minima in the curvatures as one moves from a high to a lower z-value. It is, though, by the generalization inherent in the simplified representation caused by the principles of organization and distinctiveness in long-term memory that a <u>part</u> of one phrase (or other segment) is recognized as the <u>same</u> as a <u>part</u> of another phrase (or segment). It must be this interconnection between parts of phrases (segments) which makes the compositionality in syntactic structures productive. Or, in other words: The productivity of discrete grammar appears in the moment when the compositionality inherent in the unique, non-discrete syntactic structures can be extended to apply intersegmentally as well.

In this way we can assume that a discrete grammar can be generalized from continuous syntactic structures, and become productive by means of the principles of organization and distinctiveness characteristic for the biological architecture of the human long-term memory system. The mapping from a non-discrete to a discrete grammar will thus primarily consist in the mapping from short-term to long-term memory encoding, and secondly in the productivity by the discreteness of the latter.

Bilinguals.

The model as sketched so far suggests that there may be a basic-level, non-discrete grammar which operates in short-term memory and extracts the syntactic structure from acoustic input in a perceptual process. In addition, there may be a discrete grammatical competence which functions in a much more 'logistic' manner, possibly utilizing the characteristic faculties of long-term memory in generalizing from the perceptually organized grammatical input. These divergent grammars may be radically different, or they may be ends on a continuous scale, ranging from low-level to high-level cognitive processes as the basis for the organization of input and the semantic interpretation of it. Since high-level cognitive processes operates on the output from low-level processes, these different grammatical competences can be seen as providing input for each other and exist in a feeding system. The high-level competences will typically be concept-oriented, in taking the readily processed chunks of conceptualized output from working memory as input, and should thus expectedly be operating on categorized matter. In comparison, the low-level competence should be sensory oriented.

This subdivision (of a possibly continuous scale) may be seen as receiving support from research on bilingual children. This group of language users should be particularly interesting from our viewpoint, since these children will be continuously exposed to linguistic data which possess different grammatical structures and hence to a much larger extent will have to be analytically discriminated in order to be interpretable. Thus, if the low-level competence is developed at an early stage in language acquisition and high-level grammars are developed later in a generalization process, we should in fact expect bilingual children to be in larger need of the high-level, analytically oriented grammar as compared to the unilingual children.

This seems to be the case. A number of studies have investigated the impact of bilingualism on cognitive skills, and it now seems generally approved that bilingualism favours the cognitive development. Although a large number of investigations have shown the negative impact from bilingualism on cognitive development, it seems to be the case that these are basically associated with minority language groups or other extralinguistic, social factors (Cummins & Swain 1986, p.17).

Cummins & Swain also draw a distinction between language proficiency in context-embedded communication and such proficiency in context-reduced situations. The former pertains mainly to personal, interactive communication, while the latter "relies primarily on linguistic cues to meaning and may in some cases involve suspending knowledge of the 'real' world in order to interpret (or manipulate) the logic of the communication appropriately" (p.152f). The distinction is a revision and elaboration of an earlier theoretical distinction between basic interpersonal communicative skills (BICS) and cognitive/academic language proficiency (CALP) (Cummins 1980). Although Cummins & Swain stress the amount of shared reality as the most important difference, the most interesting difference (for our purposes) between these two propsed proficiencies lies in the presence/absence of direct acoustic input in the communication: The former relies on the acoustic coding of the linguistic structure, while the latter primarily involves non-acoustic coding such as written language. What is of interest to us is that these seem to be two different competences in language users, with a different development under various conditions, particularly investigated for bilingual children.

Of even more interest are the experiments carried out on bilingual children's orientation to language. Cummins & Swain (1986, p.20) remarks: "Ianco-Worrall (1972), in a study conducted in South Africa, reported that bilingual children brought up in a one-person, one-language home environment were more oriented to the semantic rather than the acoustic properties of words and were more aware of the arbitrary assignment of words to referents than were unilingual children. [...] Ben-Zeev (1977a,1977b) has reported findings which suggest that bilinguals develop a more analytic orientation to language and more sensitivity to feedback cues. Ben-Zeev (1977b) hypothesized that bilinguals develop this analytic strategy towards language as a means of overcoming interlingual interference". Further investigations are reported which suggest that "early bilingualism can accelerate the separation of sound and meaning".

It was argued in chapter 1 of the present study that the discreteness of grammar is closely connected with the extent of (assumed or perceived) arbitrarity in the linguistic sign. In comparison, a non-discrete grammar will assume a much smaller degree of arbitrarity and a more immediate presentation of meaning in sound. This seems to be firmly supported by these findings.

Our hypothesis that bilingual children will be more oriented towards a discrete, long-term memory governed grammatical interpretation of speech is thus furthermore supported by experiments carried out and reported by Cummins & Swain (1986). A group of bilingual as well as a control group of unilingual children were tested by means of questions of the following sort: "Suppose you were making up names for things, could you then call the sun "the moon" and the moon "the sun"?" (p.24). It turned out that almost 70% of the bilingual children would accept such interchange, while only 27.5% would do so among the unilinguals. Although Cummins & Swain stress that caution is required in interpreting the results, the findings still point in the same direction as Ianco-Worrall's conclusions, that bilinguals to a larger extent than unilinguals conceive the relationship between meaning and form as arbitrary.

Why should bilinguals prefer the discrete grammar, which implies such arbitrarity in the signification? The reason must be found in a larger difficulty in communicating through a non-discrete grammar which functions by means of a direct perceptual interpretation. This could point to a function of non-discrete grammar as providing a more immediate interpretation, which, possibly due to the very short time spans allowed for such processing in short-term memory, does not give room for considerations on which perceptual code is required. (In true bilingualism, we must suppose that the two codes have radically different distributional properties, while this need not be the case for more closely related dialects or idiolects). This again could suggest that the perceptual code is acquired by learning, and must be built, possibly interactively with the updating of the database over sound relationships, in the process of language acquisition. We return to this below, in the summary of requirements for the language acquisition process.

The immediateness of perceptual interpretation which we have in mind can be illustrated by the immediateness of visual perception. When we see an ordinary object, say, a table or a chair, we normally interpret it immediately, with no delay by doubt. If, however, we encounter situations in which we frequently discover that our first and immediate interpretation was illusory or erroneous, we will learn to doubt the immediate perception and introduce reflection and assessment of the interpretations. The more often we are mistaken or misled by our own perceptual structuring and conceptualization of the incoming data, the more will we establish a hesitant and analytical assessment of our interpretations. This is exactly the situation which bilingual children face in their everyday life, if the incoming linguistic data are sometimes in one code and sometimes in another. This may be the simple reason why bilinguals tend to develop a more analytical orientation towards language.

Semantics.

The illustration can be taken even further into a simple semantic model. When we do not doubt our visual perceptual organization and conceptualization, we normally do not consider the possibility that the incoming visual data are not identical with the object which we recognize them as. On the contrary, these data are the very object: We have to perform a fairly complex philosophical analysis to be able to distinguish between these incoming data and the 'Ding an sich'. This immediateness, this undoubted identification of the perceived data with the reality we ascribe to them, is closely parallelled in a non-arbitrary conception of the signification in speech. It is not irrelevant which form these visual sensory data have, since we do not distinguish between the data and the object: If the visual data are different, so is their 'referent' (the object) as well. For a semantics which works along the same principles in a perceptual process, the identification of a meaning in the perceptually structured incoming acoustic data will be equally undoubted and immediate. The semantics of non-discrete grammar will consist in assignment of meaning to the distributional structure (in our investigation represented by a curving surface in three-dimensional space).

The illustration of semantics in acoustics with the semantics in visual perception should not be understood too narrowly. The senses may well have different functions as to what ontological status we assign to the perceived structures, and a simple linguistic semantic model could e.g. assign priority to the ontological status of non-speech-like sensory data. In the perception of distributional structure in speech (which presumably is much more prominent and regular than the distributional structure in other auditory perceptions), this simple semantics could see an identification of the speech-induced perceptual structure with other simultaneous sensory data as the basic semantic function. In particular, we keep the possibility open for this identification to include the awareness of mental content. A defined meaning would thus be the identification of the mental content arising from the definition with the expression to be defined. (When, e.g., a childs asks for the meaning of a word). From a non-discrete point of view, the whole structure (the whole curving surface in the model we have sketched) can be assigned meaning in its identification with the totality of other simultaneous sensory data: In a completely non-discrete grammar, it will not be subdivided into smaller parts. From a syntactic point of view, this corresponds to a syntax which lists all utterances unsegmented, and from the recently mentioned language orientation view, this will be a minimally arbitrary signification: The identification is immediate and unanalyzed. Clearly, each such structure will be unique (since all utterances are unique in a non-discrete grammar with a high resolution) and have a unique meaning, which means that the interpretation will be a continuous naming function with no repetitions. In principle, this also means that since each identification (or the continuous process) is not repeated and need therefore not be recognized, it need not be stored in memory for a longer time. Thus, there is no need for a retrieval of the structure for a recognition process, and the semantics of a completely non-discrete grammar will not consist in the recognition of established meanings or meaning elements, but rather be the very identification function itself.

A first segmentation will, though, appear in the moment a part of the structure is identified with a part of the accompanying sensory data, e.g., a prominent curvature in the curving surface is identified with some other, say, visual perceptual prominent part-structure (a face, a body, some object). This falling-apart of the total structure into subparts may then be of the same form as the recognition of one part of a segment as identical to a part of another segment in the generalization process of discrete grammar. Thus the very generalization process which makes discrete grammar productive may induce even the semantic recognition process: A part of the perceived speech structure is no longer unique, but it now names a part of another perceptual structure in the same manner as it has done before. There is a repetition or reinforcement of the identification or naming by the retrieval of an identical trace from long-term memory. This would be the first step towards a more extensive segmentation and a meaning assignment of parts of the speech-structure identified with parts of other sensory structures (or conceptualizations) in a repetitive process in long-term memory. This would thus be the building of a lexicon through a broader or finer segmentation of the utterances, and a concomitant assignment of meaning to these generalized lexical entries.

The introduction of arbitrarity could be seen as the very loss of uniqueness in the meaning assignment: As long as meaning is a pure continuous naming process which is unique in each moment, the identification may be as total as our normal identification of the incoming visual sensations with the objects we assume to perceive. If more than one element from a category can represent the category equally well, and they represent this category and not themselves, then the naming is not unique, since any element from a category can name the referent equally well. This is basically the same as bilingual children will experience, when very different words (from different languages) apply to the same referents. Arbitrarity may thus be the product of a simple alienization process by the segmentation and the categorization inherent in the long-term memory storage format. In linguistic terms, it can be interpreted as a separation of sound, meaning and referent.

This model can of course not account for <u>why</u> this identification should fall apart and thus why a discrete grammar with arbitrary linguistic signs should appear, but it seems a possible extension of the above findings. The proposal of a priority to non-speech-like perceptual structures is in fact the only element which we have added, but even this can be seen as supported by in particular Baddeley's findings of a special store for speech-based information and another for visual/spatial information in working memory. This does of course not tell us anything about their internal priorities, but it tells us that they may have different statuses in the perceptual process.

Language change.

We have mainly considered speech perception processes, notably from the point of view of language acquisition and the interrelationship between non-discrete and discrete grammar. We have suggested that the latter can be extracted from the former in a generalization process to yield a basic discrete grammar, possibly in the form of a phrase structure grammar. We have also suggested that discrete grammar can achieve a smaller or larger independence of the non-discrete grammar, such that logical operations can be performed on it to produce a transformational grammar as a high-level, possibly autonomous syntax.

As concerns langauge change, a question of considerable interest is which grammar becomes productive. Evidently, since the syntactic structures of (a maximally) non-discrete grammar are unique, the only truly generative capacity of a continuous syntax we can imagine are in the form of utterances which conform to the distributional properties of the sounds such as they are represented in the mental database.

Wickelgren (1976) discusses the syntax of phonetic segments, and remarks: "Speech does not consist of context-free segments, nor is it temporally segmentable" (p.247). "The smallest size units that can be cut and splices from recorded utterances to produce intelligible speech are roughly a half syllable in length, and these half syllables must mesh properly in order to produce intelligible speech" (p.249). He goes on to suggest that, from a phonetic point of view, there is so much information about the environment in each phone (corresponding to a phoneme in a phonemic analysis) due to coarticulation phenomena and perceptual cues, that speech sounds at an acoustic level should not be represented as phonemes (although such discrete elements may well exist at a higher level), but are more properly represented as context-sensitive allophones. More specificly, he considers the possibility that speech is analyzed as composed of triphonic elements, which consists of a 'kernel sound' plus the information about preceding and succeeding sounds. "Even if one were to take rather long phrases consisting of many words and scramble their context-sensitive allophonic segments, it will almost always be possible to reorder the symbols to form a unique reconstruction of the ordering of the allophones to form words in the phrase". "Such a context-sensitive coding would be said to 'cross' word boundaries". "It came as a considerable surprise to me to realize how much of the information concerning the ordering of very, very long sets of elements can be communicated by this type of extremely local information concerning the relative order of adjacent elements" (p.250f).

This notion of the representation of sounds as context-sensitive allophones receives support from cognitive psychology as well. The principle of encoding specificity has been put forward by Endel Tulving (in e.g. Wiseman and Tulving 1976 and Tulving 1979), and consists in the proposal that items are encoded with respect to the context in which they appear. Eysenck and Keane (1976) sums up the current views on context-sensitive encoding: "[...] there is now strong evidence that both recall and recognition memory are affected greatly by the similarity of context at learning and at test [i.e., at retrieval]" (p.162).

Evidently, from the point of view of a generative grammar, triphonic elements do not contain sufficient information to generate only wellformed strings, but the database extracted from the perceptual parsing of speech will contain sequential constraints over a much larger area. The output from a generative grammar which takes the information from the long-term memory database as input will probably be wellformed to a considerable extent: It will in any case be wellformed from the point of view of a non-discrete grammar. It remains to be tested experimentally to what extent the output from such a grammar can approach wellformedness from a discrete grammatical point of view.

This is when no account is taken of the possibly constraining effects from a semantic wellformedness criterion: Since non-discrete grammar performance is typically context-embedded (in the sense of Cummins & Swain (1986), i.e., pragmatically constrained), it is reasonable to assume that the generative capacity of a non-discrete grammar will be constrained not only by the syntactic information from the database, but also by the semantic function. We have suggested that the basic semantic function in non-discrete grammar is the naming function, which means that the output of a generative grammar may also be constrained by the requirement that an appropriate identification of the generated sound sequence with the additional perceptual data must be possible. That is, the grammar will be pragmatically constrained as well. A methodological problem in this connection is, though, how the pragmatic appropriateness is measured. To evaluate the appropriateness of a semantic identification in a non-communicational situation a certain degree of experience is required: The output of previous semantic performance (in the sense of the match between sounds and pragmatic setting) must somehow be stored in retrievable form in long-term memory and constitute the basis for the wellformedness evaluation. In a communicational situation, though, the wellformedness will be assessable by the response from the environment, as a learning process. To the extent that the output from previous performance must be accessed as a constraining factor, the wellformedness of a generative non-discrete grammar will rely on information from long-term memory. We must therefore assume that the semantic constraining of non-discrete generative grammar is gradually achieved in the learning process of language acquisition. (Typically, the learning of these constraints may be concomitant with the generalization process in which a discrete grammatical competence is built).

We can thus assume that even a non-discrete grammar can form the basis for the production of wellformed utterances. Evidently, since the utterances are generated from the information in the database, they will conform perfectly to the syntactic structures of non-discrete grammatical competence. There will be no updating of the database. Non-discrete speech generation will not cause language change.

The situation is, though, different for the discrete grammar. A generative discrete grammar based on generalized rules for the compositionality of segments will create novel utterances in a more profound manner: By its discreteness, the sound co-occurrences across segment borders will not necessarily conform to the conditional probabilities in the database. Moreover, if, in addition, articulation is based on the discrete competence (although it may seem more reasonable to assume that articulation is context-sensitive in the sense of Wickelgren), the deviations from the database probabilities may be even more dramatic.

The output from a generative discrete grammar may therefore deviate drastically from the output of a non-discrete grammar. If general rules relate sufficiently large categories of lexical entries, the output from a discrete grammar may fail to be interpretable in a non-discrete grammar. Such language production may of course exist under the assumption that it will be interpreted in a discrete competence, but for natural oral language, we must assume that the generative capacity of discrete grammar is constrained by the requirement of interpretability in non-dicrete grammar. The assumed universal phenomenon of motherese or baby talk is a simple example of an extreme constraining of the output of a grammar.

If, therefore, child grammar develops from non-discrete to discrete productivity, and the latter must approximate the structures of the former in order to obtain wellformedness of the utterances (i.e., the produced utterance must contain an interpretable perceptual structure when parsed in short-term memory), then we will find that each new generation will generate an approximation to the speech structures of the former. If the productive language is discrete and rule-governed, then a slight deviation from a target non-discrete structure is unavoidable in all utterances, since the discrete units have a broader resolution, the rules are general and the generated structure consequently is not unique. Within this model, therefore, language change is inevitable by the very discreteness of productive grammar, but, at the same time, the change from generation to generation is bound to be small, by the constraint that the produced utterance must be interpretable in a short-term memory perceptual parsing.

This feedback system, in which discrete grammar is a generalization over the non-discrete grammar, and, by the generative capacity of discrete grammar, new non-discrete grammatical structures are created and form the basis for revised generalizations in discrete grammar, constitutes a principled explanation for language change. The basis for this feedback model must be found in the isomorphism but not identity between the distributional structures of the output of the two grammars.

As to the degree of the isomorphism which we reported above, it must be emphasized that the full syntactic structure extracted from the distributional properties of the sounds is probably much more rich and complex than the approximations to binary branching structures which we presented in the xz-diagrams. In particular, the y-values may give additional cues to the syntactic interpretation. Therefore, the simple isomorphism which we have pointed to is only a part of what can be the basis for such a generalization (including categorization and rule-formation) from non-discrete to discrete structure, and the precise mechanisms in this process can be more complex than we have sketched here. The poverty (and non-acoustic basis) of our data do, though, not allow for any further investigations of this process.

We may assume a number of versions of this model. The generative capacity of a grammar can be constrained by requirements on syntactic wellformedness only, or semantic interpretability may come in addition. Also, the constraints may vary as to the assumed degree of discreteness in the competence of the adressee. These versions need not be mutually exclusive. It is fully possible that there are more constraints on the generative capacity of discrete grammar in what Cummins & Swain (1986) refer to as a context-embedded communicational situation than there is in the context-reduced communication. This may have to do with which memory system is the basis for the semantic interpretation. nally, tIf a child (or an adult) interprets the utterance by means of non-discrete grammar, then it is reasonable to assume that in order to be communicationally successful, the generative capacity of the discrete grammar must be constrained to yield only utterances which can be interpreted by means of non-discrete grammar, while this constraint will not be present to the same extent if the adressee can utilize the discrete grammatical competence in the semantic interpretation of the utterance (or syntagm), for example while reading a text.

As will have become clear, the distinction between a non-discrete and a dicrete grammar need not be fully discrete. Rather, the model suggests a continuum (to the extent that such continua is found to exist among the different memory systems as well) ranging from the absolutely non-discrete grammar, characterized by an immediate semantic interpretation implying a full identity (i.e., non-arbitrarity) between the sound and its meaning, to the absolutely discrete grammar in which the discrete lexical items are principally detached from their referents, such as we find in logical systems or in artificial (e.g. computer) languages.

The rate of language change will thus depend on the degree of constraint on the generative capacity of discrete grammar. A minimally constrained grammar will generate large gaps between the perceptually conditioned syntactic structures of its phonetic realization and the syntax of the non-dicrete grammar which it is generalized from. It will cause an extensive updating of the database of sound probabilities through the non-discrete parsing of its output, and consequently a rapid language change.

If we assume that the degree of discreteness of the grammar is culturally conditioned, in the sense of being enhanced by literacy (which is a highly discrete analysis of language), schooling and a cultural favouring of rationalistic (logistic) thought, then we should expect to find that the rate of language change is culturally conditioned as well, since a non-discrete grammar will cause a minimal language change, while a fully discrete grammatical competence can allow for very rapid change. This can to some extent be seen as supported by the notable slow pace in the language change in some illiterate cultures, such as e.g. Polynesian and Australian languages (Even Hovdhaugen, personal communication) compared to the notably rapid change in Indo-European languages. The hypothesis is to some extent testable, if an anthropologically defined measure on cultural conditioning of dicreteness can be set up and correlated with the degree of diversity among related languages and the time span and extent of their separation.

If we consider more closely the direct impact on the continuous syntactic structures (in the form which we have suggested above) from the generativity of discrete grammar, we will find that e.g. the productivity of phrase structure rules (or, more precisely, _segment_ structure rules, which comprises morphological rules as well) will tend to dissolve dependencies across segment borders. In the xy-diagrams, i.e., the curves for the density function (for a certain z-value) above, this will appear as the reinforcement of a local minimum or the insertion of a new local minimum over what was previously one curvature. In the xz-diagrams (which we have suggested mirror the discrete binary branching structures), this will appear either by the extension upwards for a branch, i.e., the node of attachment will appear higher up in the diagram, or it will appear as a completely new branch. In either case, the element which has gained productivity will - in a discrete grammatical generalization - attain a _larger syntactic scope_. If for example a suffix with a separate branch becomes more productive, its node of attachment leftwards (i.e., to the stem) will most probably move upwards by the decreased cross-segmental dependency in the overall distribution of sounds. (It must,though, be emphasized that this is the _general_ tendency: It may of course well be counterbalanced by other processes in other parts of the distribution of sounds).

This conforms well to the general tendency for inflectional morphemes (defined by, amongst other things, a larger productivity than derivational morphemes) to appear peripherily and to have a larger syntactic scope than derivational morphemes. Bybee (1985) reports from a cross-linguistic investigation of unrelated languages a strong tendency for _increased morphophonological fusion_ across morpheme boundaries closer to the root than further away from it. This amounts to a general tendency in the historical development of languages for boundary degrading closer to the root. To the extent that such fusion is conditioned by general phonological rules or regularities (which we must assume it to be in the vast majority of cases), these processes will, by the introduction of cross-segmental dependencies, tend to straighten out the xy-curves in the area where the dependencies are introduced. In the xz-diagrams, this will appear as a _lowering_ of the node from which the suffix branches extend. The general tendency for increasing morphophonological fusion towards the root will thus be reflected in a general tendency for leftbranching word-internal syntactic trees in the non-discrete syntactic structures (that is, in their xz-representation) in suffixed forms and rightbranching in prefixed forms. This is well in line with the tendency for suffixing languages to have a leftbranching word-internal structure (such as we typically find it in Hungarian, Turkish, West Greenlandic, Tamil etc.).

Finally, the changes in the distribution of sounds which the generativity of discrete grammar induces need not be reflected in neither upgrading nor downgrading of segment boundaries, but may result in slight displacements in the curvatures of the non-discrete structures only. By the categorization principles of long-term memory encoding, such displacements in non-discrete syntax may remain unnoticed in the generalized discrete grammar untill a certain point in the historical development, where a limit for difference (dependent on the principles for cognitive organization) is transgressed and a resegmentation (compared to earlier segmentation) is performed.

These three processes - boundary upgrading and downgrading as well as resegmentation - are thus natural consequences of the feedback system. The fundamental point in the present context is that these restructurings stem directly from the following two principles:

1. dicrete grammar is generalized from non-discrete grammar
2. by its generalized form, discrete grammar is bound to induce changes in the distribution of sounds

The rate and extent of language change will thus within this model ultimately depend on the distance between the two grammars. And since discrete grammar induces changes in non-discrete grammar, the discrete grammars of successive genearations are bound to be different.

Language acquisition and the innateness hypothesis.

As discussed above, the present model suggests an alternative to the innateness hypothesis, since it can account for both positive and negative evidence for the setting up of a discrete grammar. More specificly, the model also proposes that a non-discrete linguistic competence is possible, and it assumes that the non-discrete grammar, at a low cognitive level, is primary, and the discrete grammar, at a higher cognitive level, is secondary. The high-level grammars are generalizations from the grammars at lower levels, and a generalized, high-level grammar will typically be in the form of a phrase structure grammar. The model suggests an increasingly autonomous syntax as the higher levels are approached, and assumes that general cognitive processes (i.e., not specificly linguistic) are responsible for the generalizations. An optional transformational component can be seen as logical operations on the generalized structures.

Language acquisition will typically consist in at least two different processes:

1. For the acquisition of the non-discrete grammar, the child must learn the sequential constraints and the full set of sound probabilities which constitutes the basis for the syntactic function. Furthermore, as the data on bilingual children may suggest, there is the possibility that the child must break the linguistic code for the syntactic function in his or her particular language. If any language - given the database of its sound co-occurrence probabilities - could be parsed by means of the same syntactic function, then it would seem reasonable that bilingual children could meet any acoustic input with the same perceptual processing. If, however, we interpret the data on bilingual children such that their particular development is motivated by the different languages requiring different processing, then we must assume that every language (or, possibly, every language type) is characterized by a particular syntactic function. Alternatively, the possibility must be kept open that the syntactic function is the same, but the parsing function (in our investigation: the density function) differs from language to language. (It may e.g. well be that prefixing and suffixing languages require different y-value assignment, and else that other typological differences can influence the generation of the non-discrete structure). If this be the case, these functions can hardly be innate, and the child must find the proper way of structuring the linguistic data.

2. When a first non-discrete grammar is mastered, a possible subsequent language acquisition process will consist in the generalization from low-level to higher-level grammars in approaching the discrete grammatical competence.

The positive evidence for discrete grammar is the structures found in non-discrete grammar. Positive evidence is also the biological design of long-term memory and the particular cognitive characteristics of the categorization process.

The negative evidence is possibly the constraining factors discussed in the above paragraphs (on language change): The discrete grammar is constrained by the requirement that the phonetic realization of its output must be interpretable in a non-discrete competence. It is also important to note that the development of the dicrete grammar will be a continuous process, in which each step forward is constrained by the previous development. There are no huge jumps which could suggest an arbitrarily designed grammar.

APPENDIX A

The following analyzed syntagms are chosen rather randomly from Sándor Márai's novel 'Egy polgár vallomásai' p.11 (syntagm 1-15) and p.15f (syntagm 16-28). The syntagms are successive sentences in the text. They are presented in a basically morphemic segmentation. There are four lines in most of the transcriptions: The first line is in traditional orthography, the second line shows the coding according to corpus A. The diagrams for corpus A expose the syntagms in this coding. (Note that a symbol is never written twice in succession on the bottom line. Thus when syntagm 30 starts with 's ha az' this is rendered as 's ha z'). The third and the fourth line contain some grammatical information and a basic approximation to the semantic interpretation. Finally, the syntagms are presented in translation.

The following abbreviations are used:

1sg, 3sg, 3pl -	grammatical person for verbs and nouns
NOM -	derives nouns from any word class
VB -	derives verbs
ADJ -	derives adjectives
ADV -	derives adverbs
PP -	derives postpositions
PRET -	preterite
PR.PT-	present participle
PRT -	perfect participle
POT -	potential form (-hat/-het)
PERF -	perfective aspect
IMP -	imperative
ACC -	accusative
DAT -	dative
NEG -	negation
PLUR -	plural

The coding for the analyses in the diagrams over corpus D is not given (but should be fairly easy to induce from the parallel diagram from corpus A). It is identical to the coding in corpus A except for the vowels, which are the following:

Stressed vowels		Unstressed vowels	
I -	[i]	i -	[i]
X -	[y]	x -	[y]
) -	[e:]	? -	[e:]
E -	[e]	e -	[e]
W -	[ø]	w -	[ø]
(-	[a:]	! -	[a:]
A -	[å]	a -	[å]
O -	[o]	o -	[o]
U -	[u]	u -	[u]

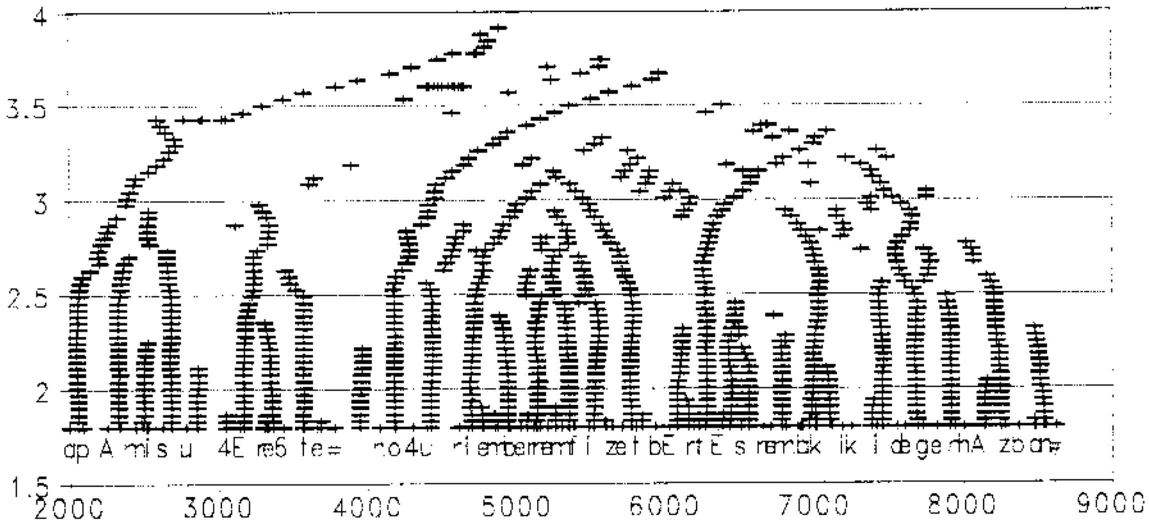
Syntagm 1

apá - m	is	úgy	érez - t - e,	hogy	úr - i -ember	nem
apA - m	is	u4	Ere6 - t - e#	ho4	ur - i -ember	nem
father-1sg	also	so	feel-PRET-3sg	that	gentleman-ADJ-man	not
father-my	also	so	felt,	that	a gentleman	doesn't

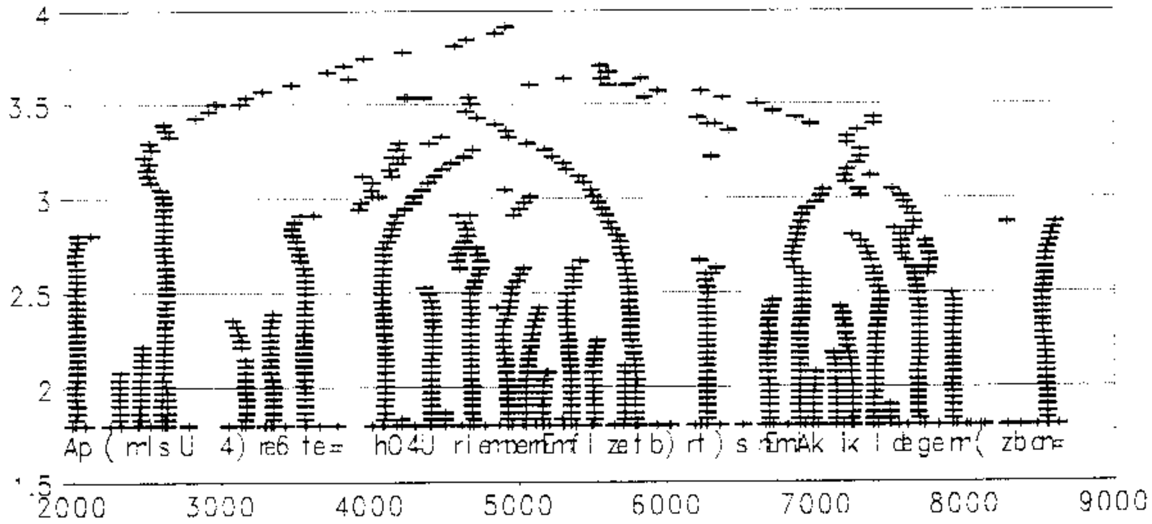
fizet	bér - t,	és	nem	lakik	idegen	ház - ban
fizet	bEr - t#	Es	nem	lakik	idegen	hAz - ban
pay	rent - ACC	and	not	live-3sg	strange	house-in
pay	rent	and	doesn't	live	other's	house-in

"My father also had the feeling, that a gentleman doesn't pay rent and doesn't live in other people's house".

Corpus A:



Corpus D:



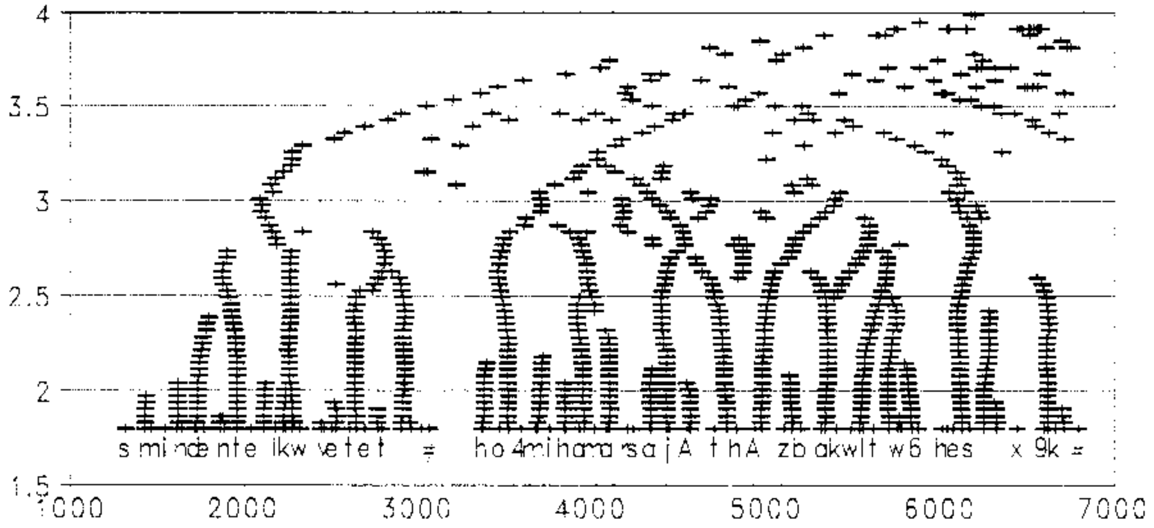
Syntagm 2

s	mind - en - t	el - követ - ett,	hogy	mi - hamar	saját
s	mind - en - t	el - kwvet - et #	ho4	mi - hamar	sajAt
and	every-NOM -ACC	PERF- follow - PRET	that	what- soon	own
and	everything-ACC	followed (he)	that	very soon	our own

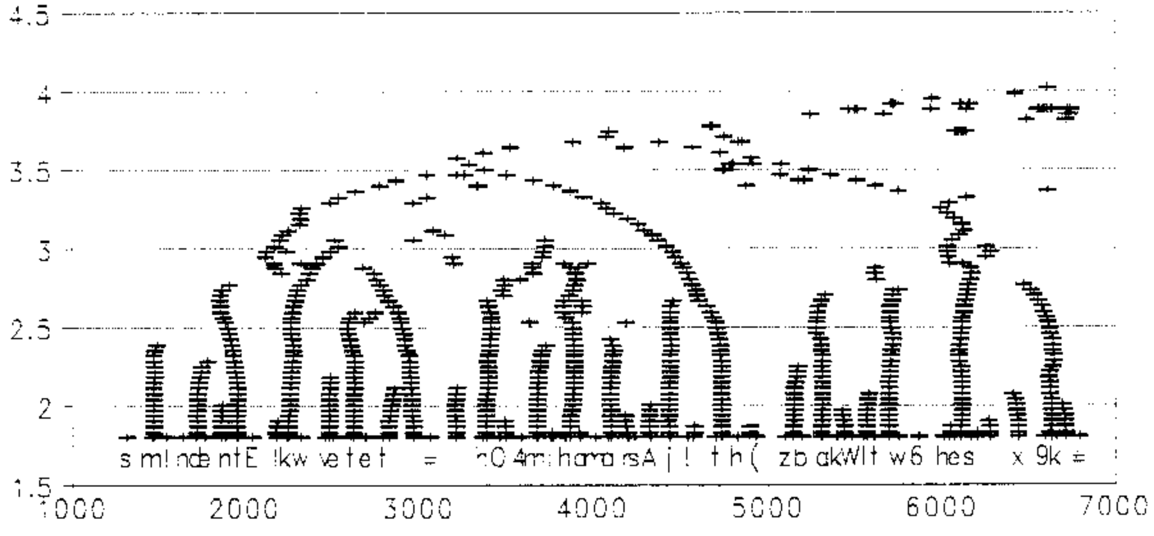
ház - ba	költöz - hes - s - ünk
hAz - ba	kwltw6 - hes - x9k
house-into	move - POT -IMP- 1.pl
house-into	(in order that) we could move

"and he set all wheels in motion (he did everything) in order for us to be able to move into our own house as soon as possible"

Corpus A:



Corpus D:



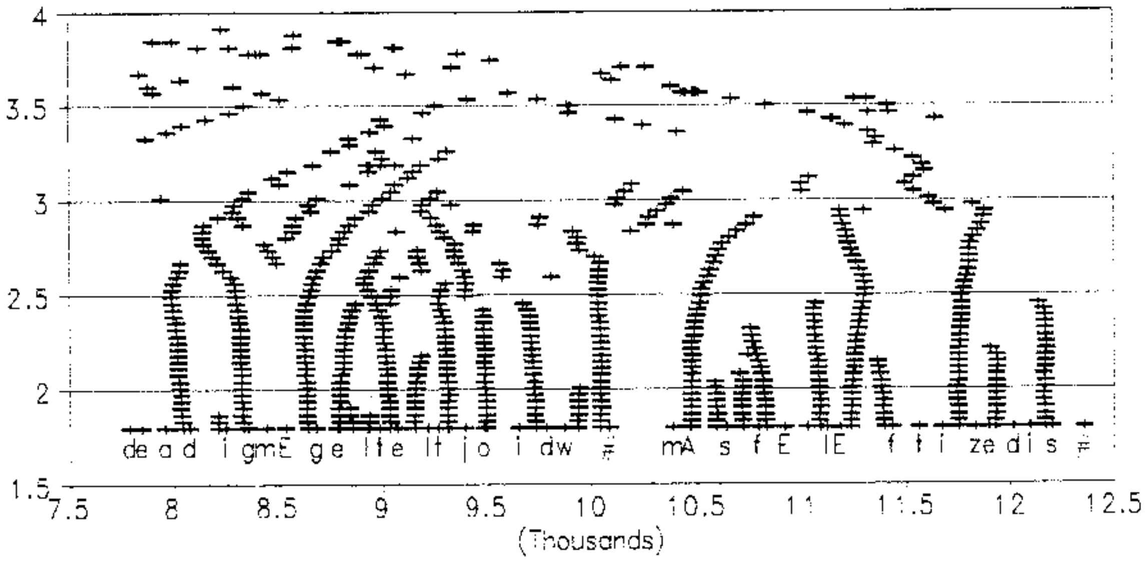
Syntagm 3

de	ad-dig	még	el - tel - t	jó	idô,	más - fél
de	ad- ig	mEg	el - tel - t	jo	idw#	mAs - fEl
but	that-to	still	PERF-fill-PRET	good	time,	other-half
but	untill	still	passed	long	time,	one-and-a-half

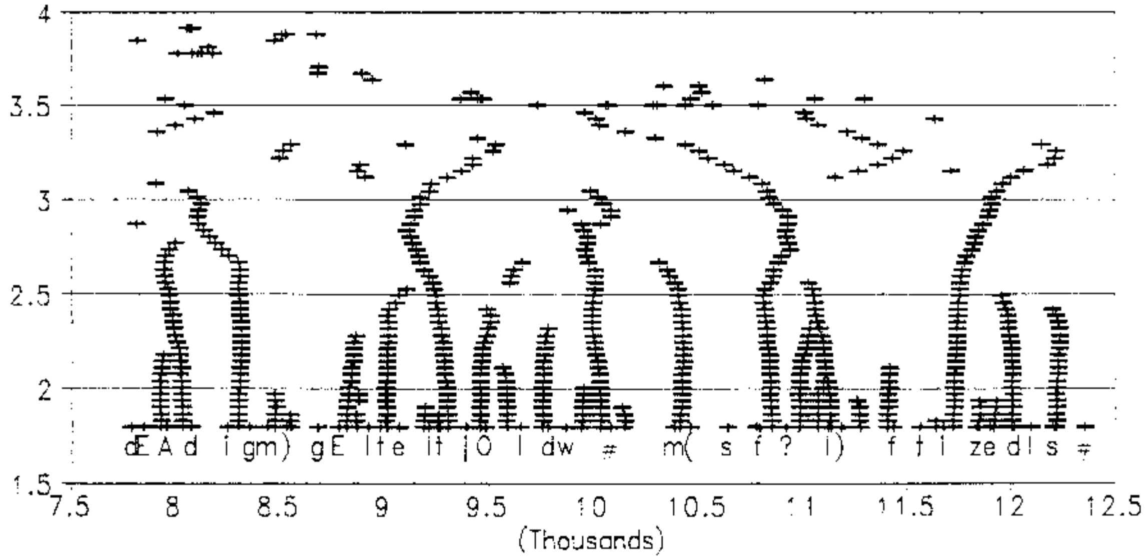
év - tiz - ed	is
Ef - tiz - ed	is
year-ten -NOM	also
decade	also

"But before this (could be), a long time went by; fifteen years, in fact".

Corpus A:



Corpus D:

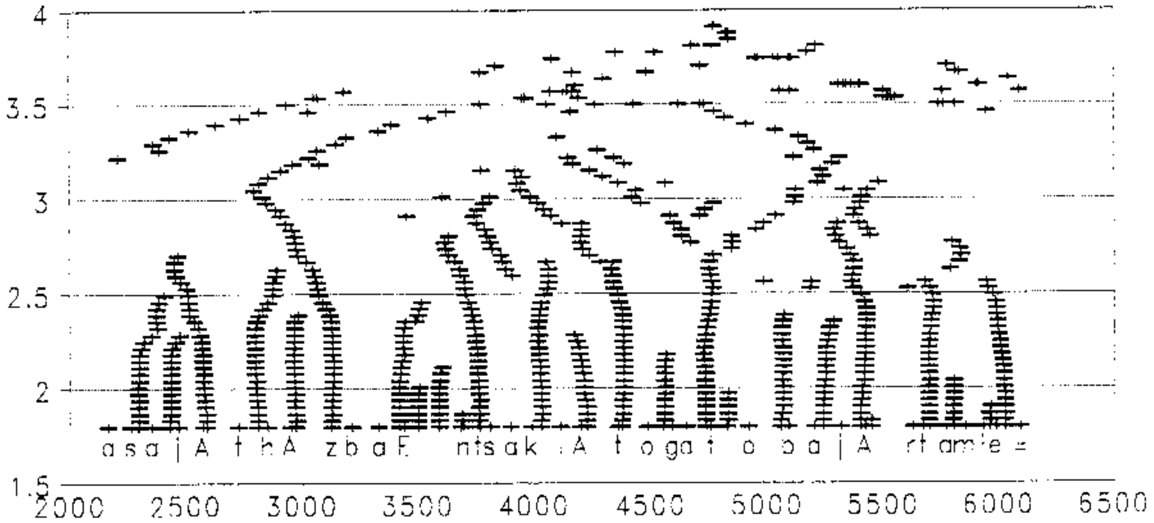


Syntagm 4

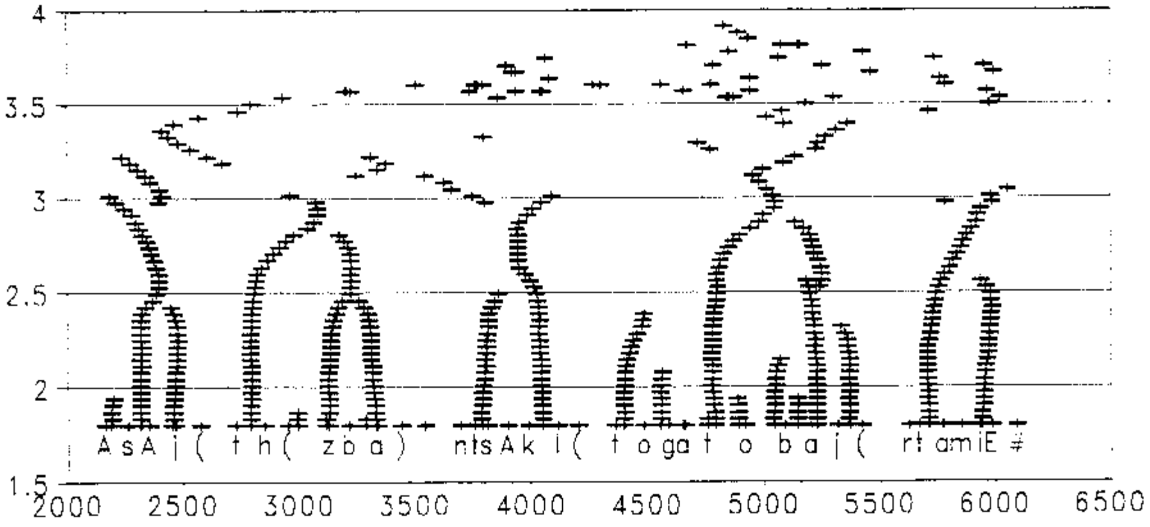
a saját ház - ba én csak lát-o-gat - ó - ba jár - t - am le
a sajAt hAz - ba En tsak lAt-o-gat - o - ba jAr - t - am le
the own house-into I only see-NOM-VB-NOM-into come-PRET-1sg down
our own house I only on visit came down

"To our own house I only came as a visitor"

Corpus A:



Corpus D:

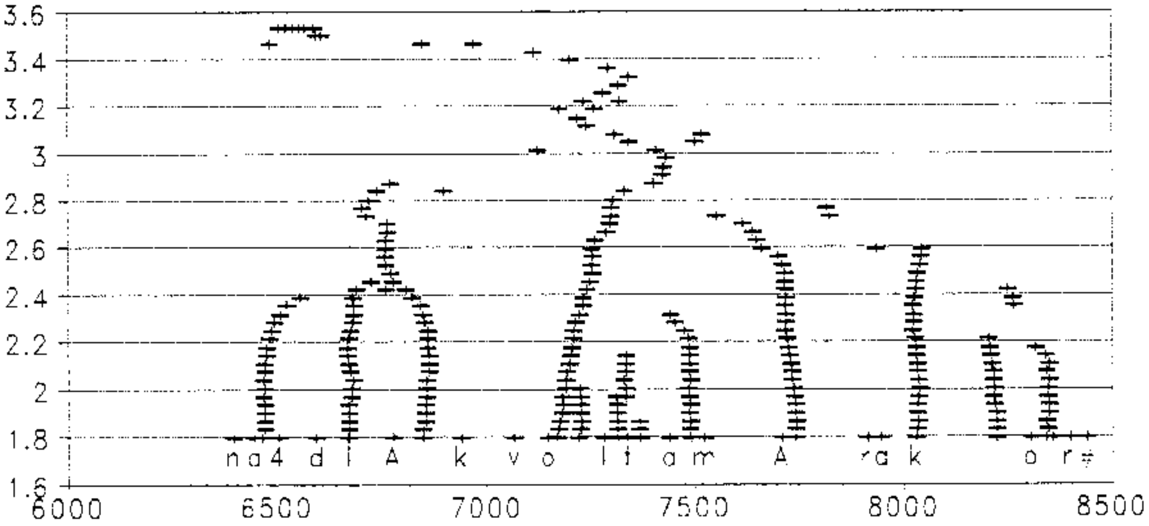


Syntagm 5

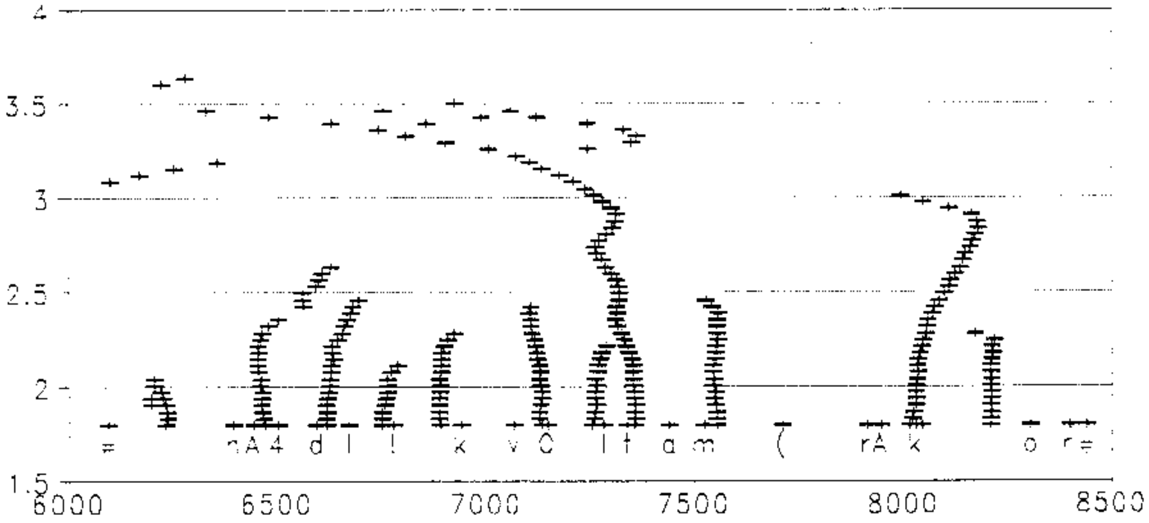
nagy	diák	vol- t - am	már	ak - kor
na4	diAk	vol- t - am	mAr	ak - or
big	pupil	be-PRET-1sg	already	that- time
big	pupil	I was	already	then

"At that time I was already a big (mature) schoolboy"

Corpus A:



Corpus D:



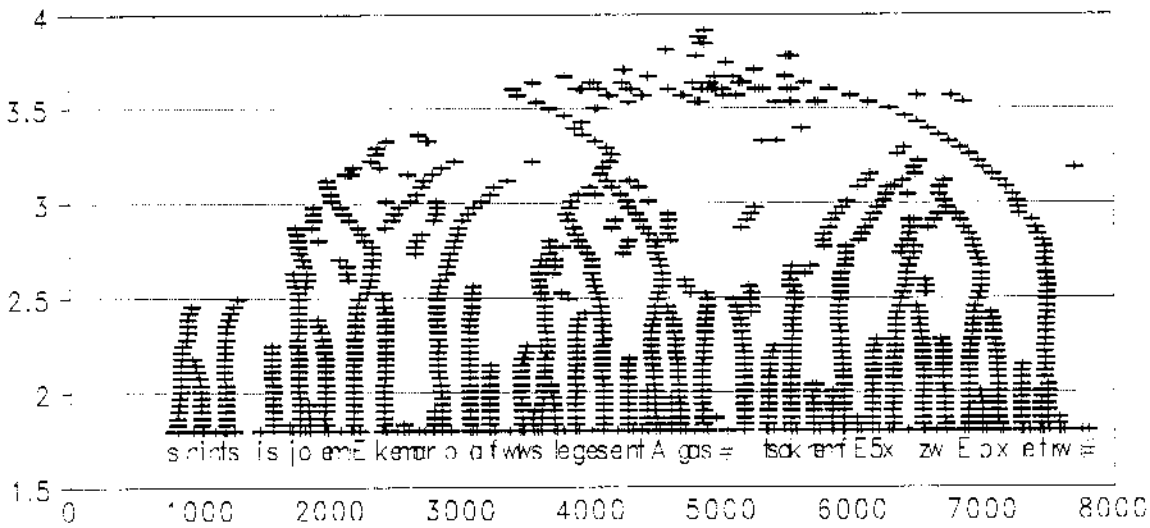
Syntagm 6

s	nincs	is	jó	emlék-em	ar - ról	a	föl-ös -
s	nints	is	jo	emlEk-em	ar - ol	a	fwl-ws -
and	be-not	too	good	memory-1sg	that-about	the	up -ADJ-
and	there	is no	good	memories	for me about	the	super -

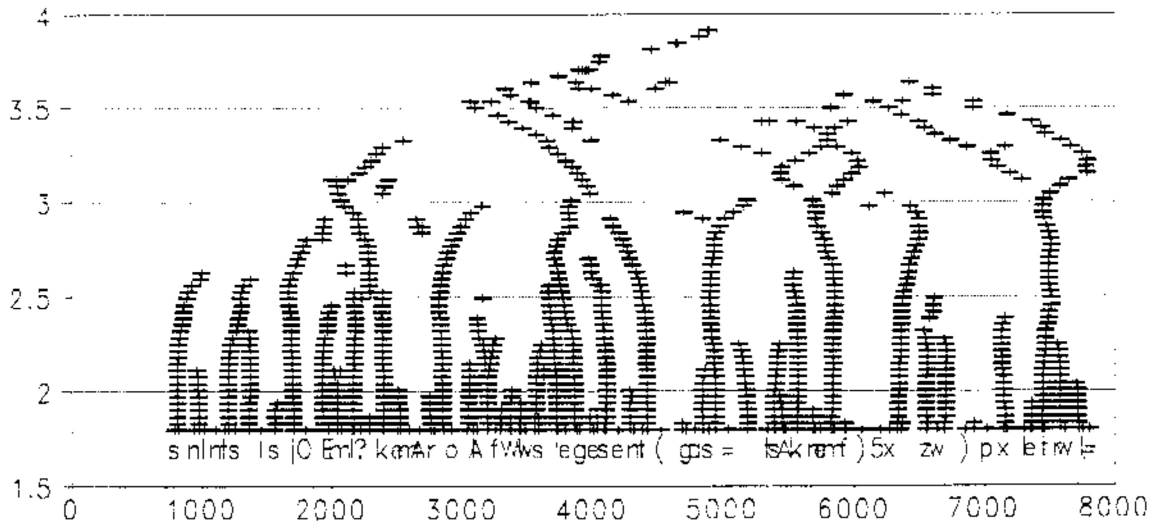
leges - en	tágas,	csak-nem	fényûzô	épül - et - rôl
leges - en	tAgas#	tsak-nem	fE5xzw	Epxl - et - rwl
ADJ -ADV	spacious	only-not	luxurious	build-NOM-about
fluously	spacious	almost	luxurious	building

"and I don't have good memories from this superfluously spacious, almost luxurious building".

Corpus A:



Corpus D:

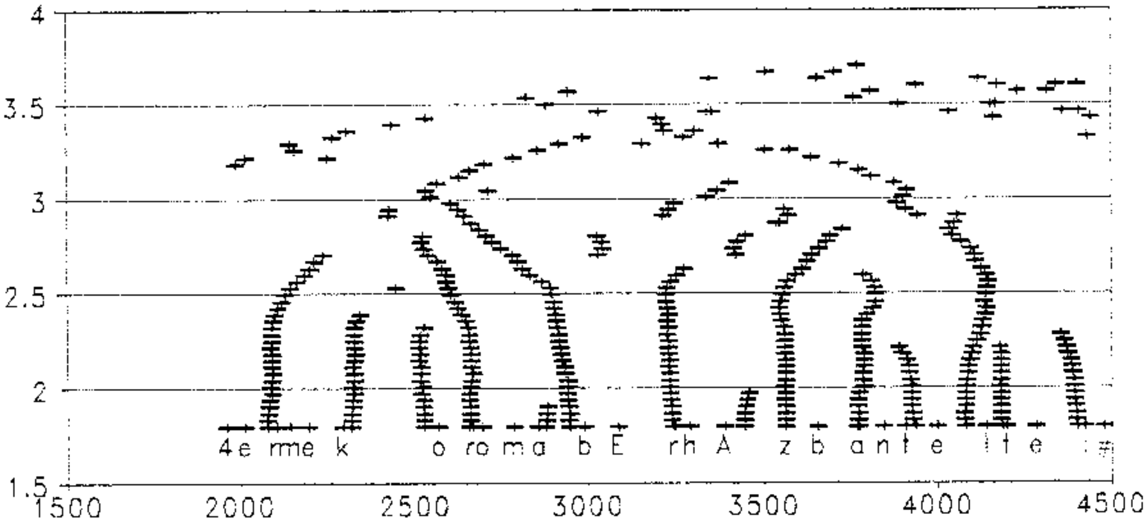


<u>Syntagm 7</u>

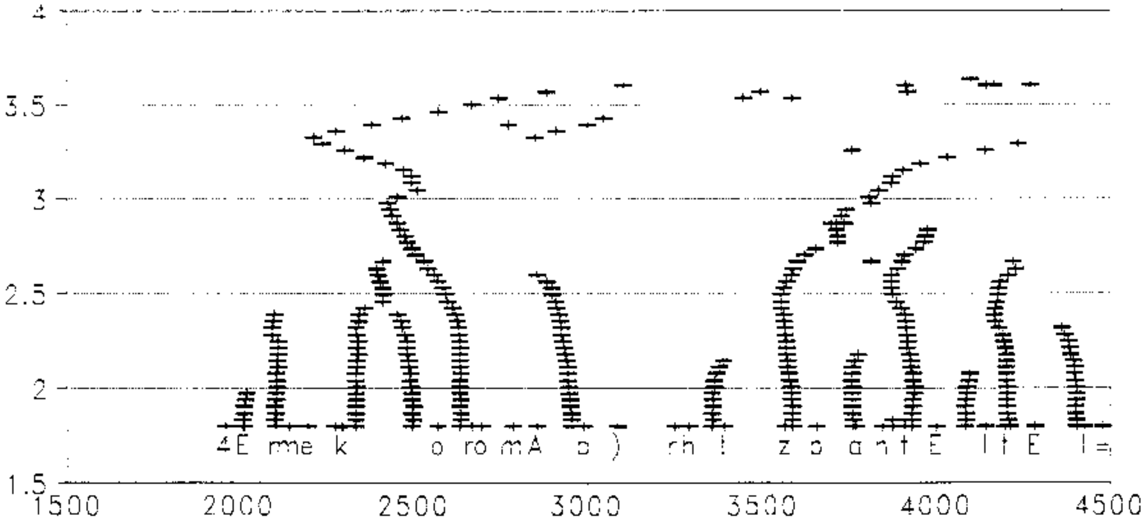
gyermek-kor-om	a	bér - ház - ban	tel-t	el
4ermek- or -om	a	bEr - hAz - ban	tel-t	el
child-time-1.sg	the	rent-house- in	fill-PRET	PERF
my childhood	the	rented house-in	went by	

"My childhood went by in a block of flats".

Corpus A:



Corpus D:



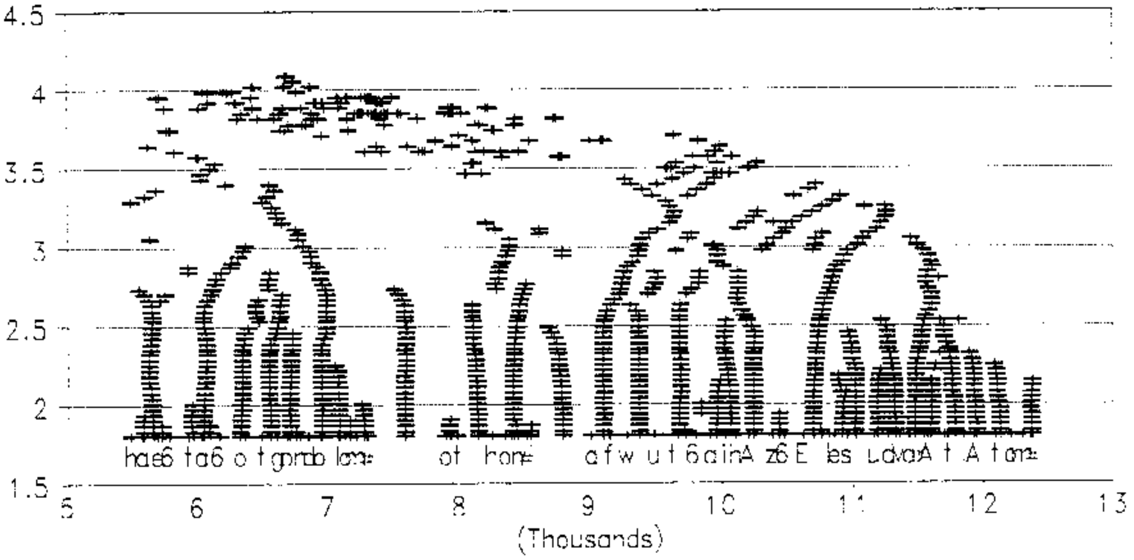
Syntagm 8

ha ez -t a szô -t gond - ol - om: ott-hon a
ha e6- t a 6o -t gond - ol - om# ot -hon# a
if this-ACC the word-ACC problem- VB -1sg there-home the
if this word I think of : "home" (then) the

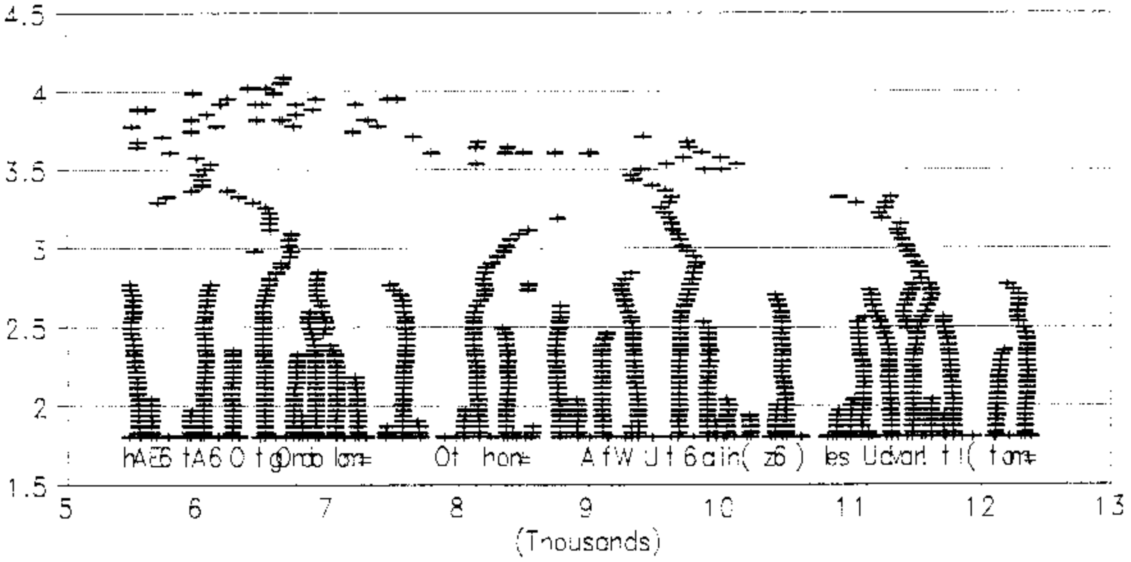
fô utca - i ház szél - es udvar - á - t lát-om
fw ut6a - i hAz 6El - es udvar - A - t lAt-om
main street-ADJ house edge-ADJ backyard-3sg-ACC see-1sg
main-street house's broad backyard I see

"If I think of this word: "home" - then I see the backyard of this main street house".

Corpus A:



Corpus D:



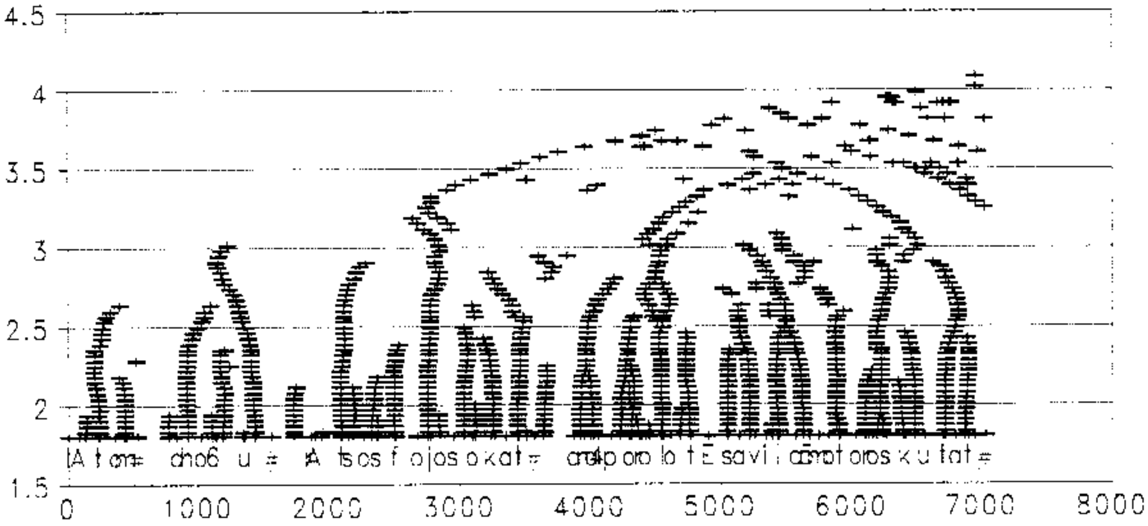
Syntagm 9

a hossz - ú, rács - os foly - os - ó - k - at, a nagy
a ho6 - u# rAts - os foj - os - o - k - at# a na4
the length-ADJ,banister- ADJ flow- ADJ -NOM -PLUR -ACC, the big
the long, with-banister corridors, the big

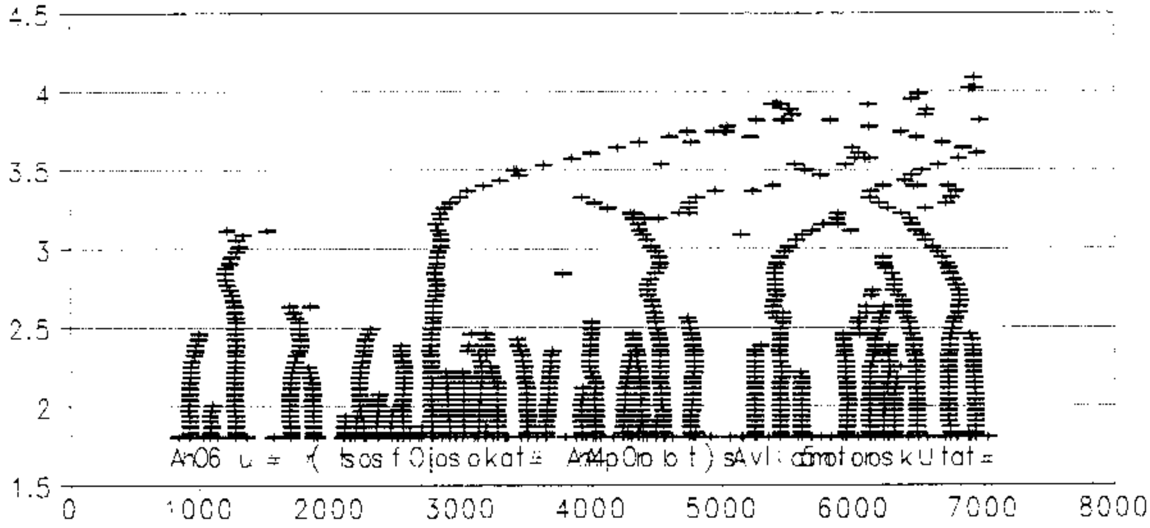
por - ol - ó - t és a villany - motor- os kut - at
por - ol - o - t Es a vila5 - motor- os kut - at
dust- VB -PR.PT-ACC and the electricity- motor-ADJ well-ACC
dust-beater and the electric-motored well

"the long, 'banistered' corridors (ACC), the big dustbeater (ACC) and the well with an electric motor (ACC)".

Corpus A:



Corpus D:



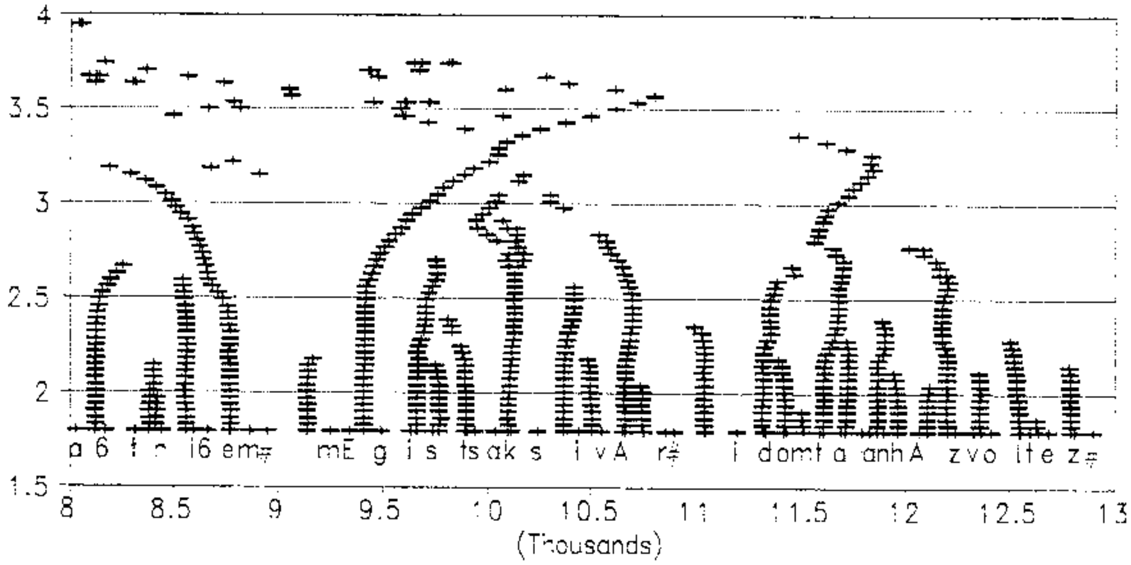
Syntagm 10

azt hisz - em, még - is - csak sivár, idom-talan
a6t hi6 - em# mEg - is - tsak sivAr# idom-talan
that believe-1sg, still-also-only dreary, form-less
that believe-I, that after all lifeless, form-less

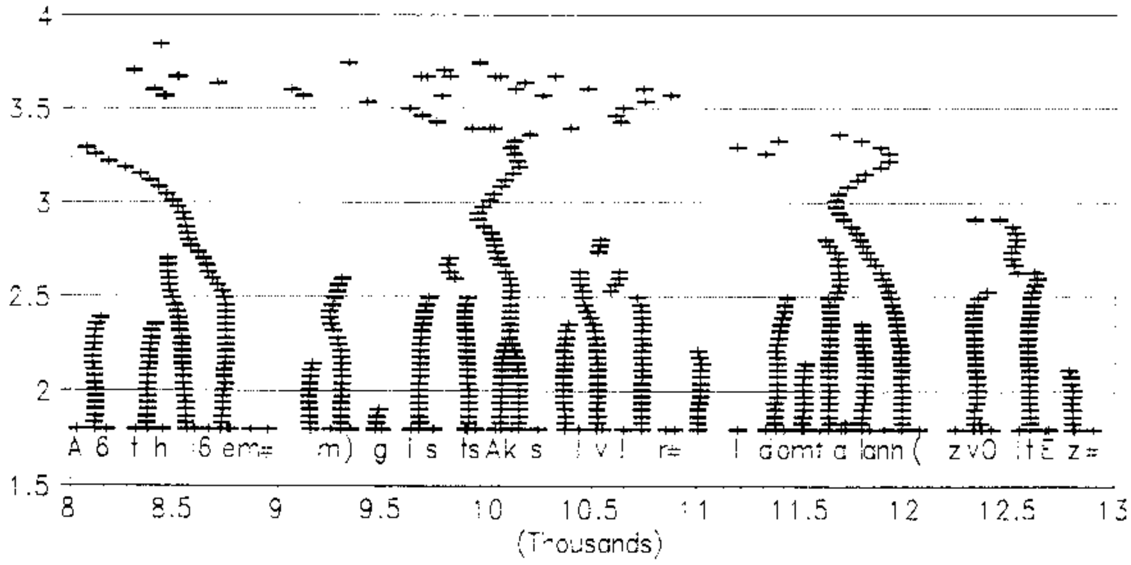
ház vol-t ez
hAz vol-t ez
house be-PRET this
house was this

"I think that, after all, this was a lifeless, formless house".

Corpus A:



Corpus D:

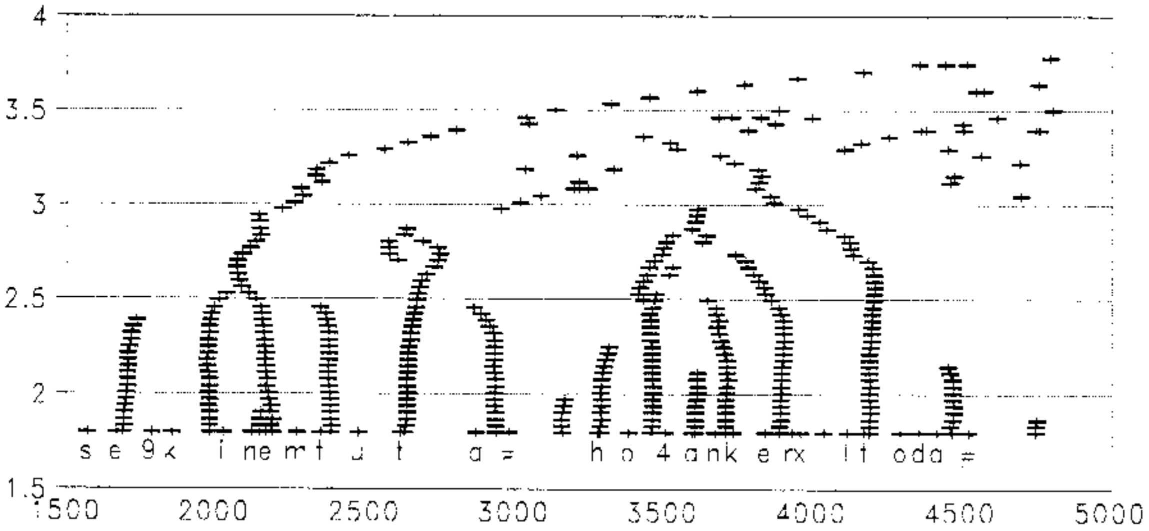


Syntagm 11

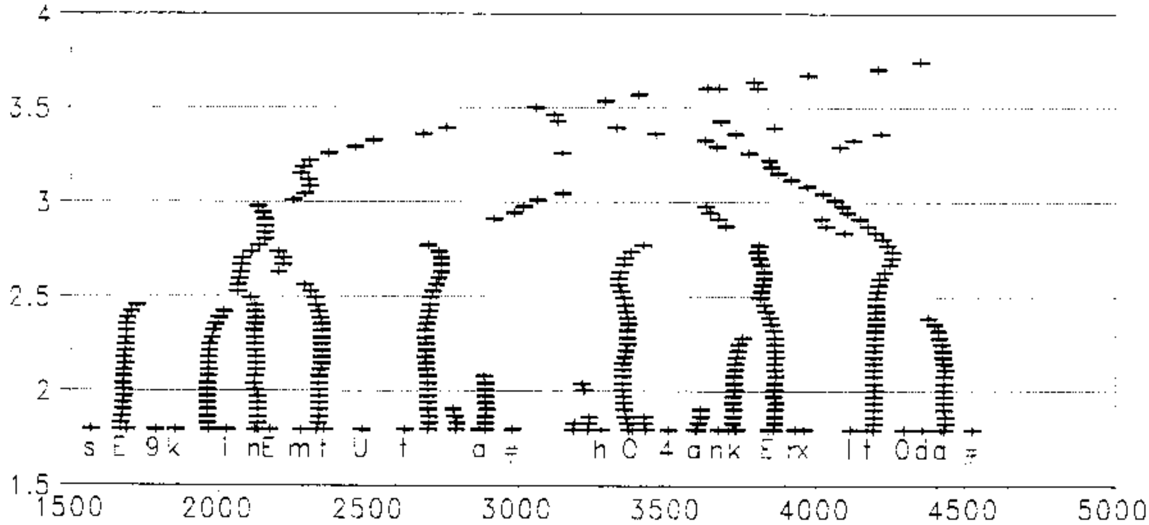
sen-ki	nem	tud - t - a,	hogy-an	kerül - t	oda
se9-ki	nem	tu - t - a#	ho4 -an	kerxl - t	oda
no-body	not	know-PRET-3sg	how-ADV	arrive-PRET	thither
nobody		knew	how	s/he came	there

"Nobody knew how s/he came there".

Corpus A:



Corpus D:



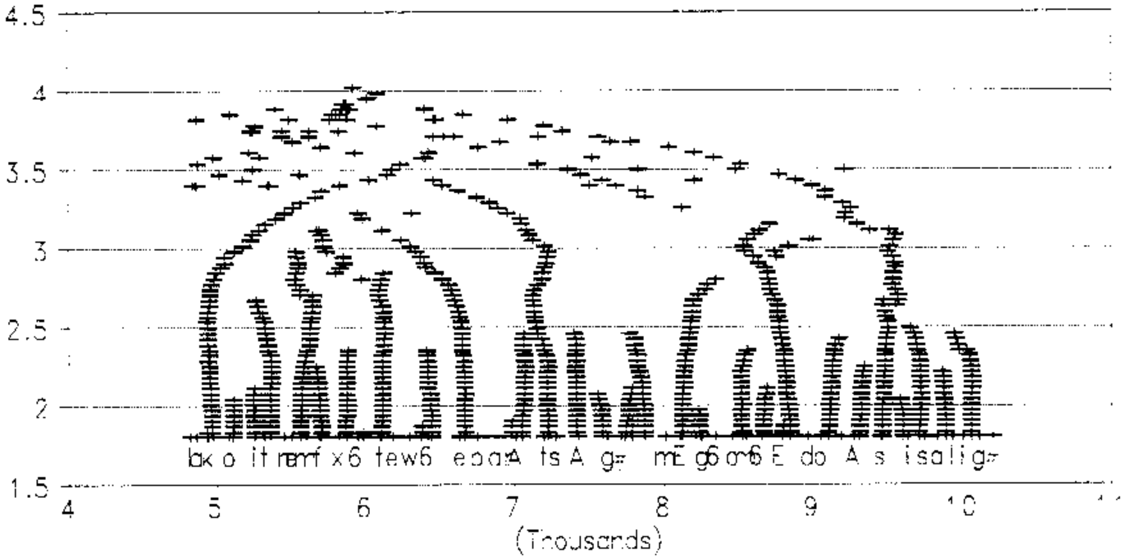
Syntagm 12

lak - ó - i - t	nem	fûz - t - e	össze	barát - ság,
lak - o - i - t	nem	fx6 - t - e	w6e	barAt - sAg#
live-PR.PT-PLUR-ACC	not	bind-PRET-3sg	together	friend-ship,

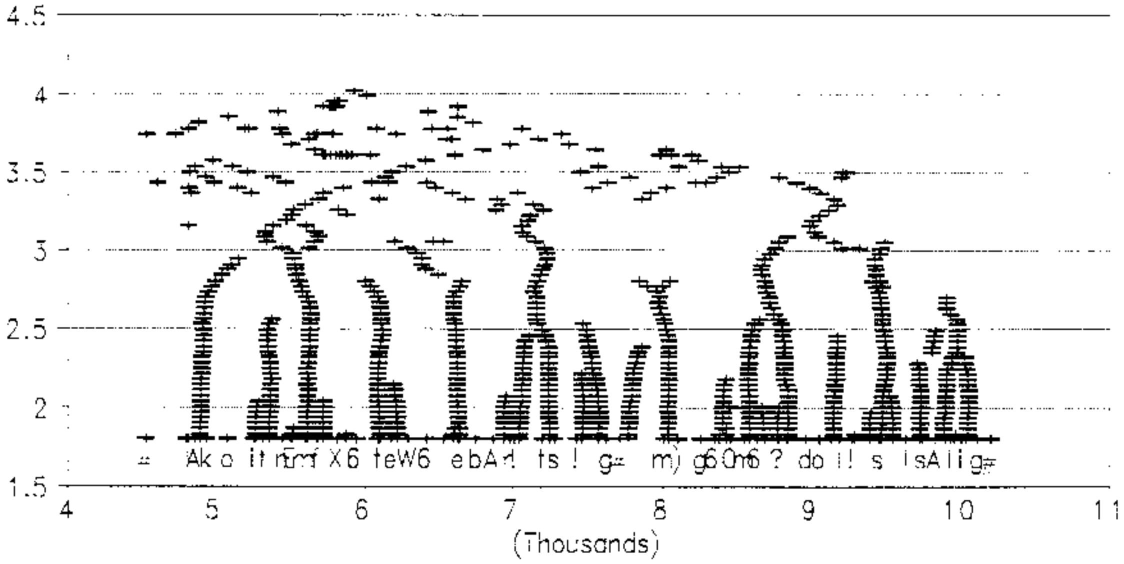
még	szomszéd -ol- ás	is	alig
mEg	6om6Ed -ol- As	is	alig
even	neighbour-VB-NOM	also	hardly

"its tenants were not bound together in friendship, not even in 'neighbourship'".

Corpus A:



Corpus D:



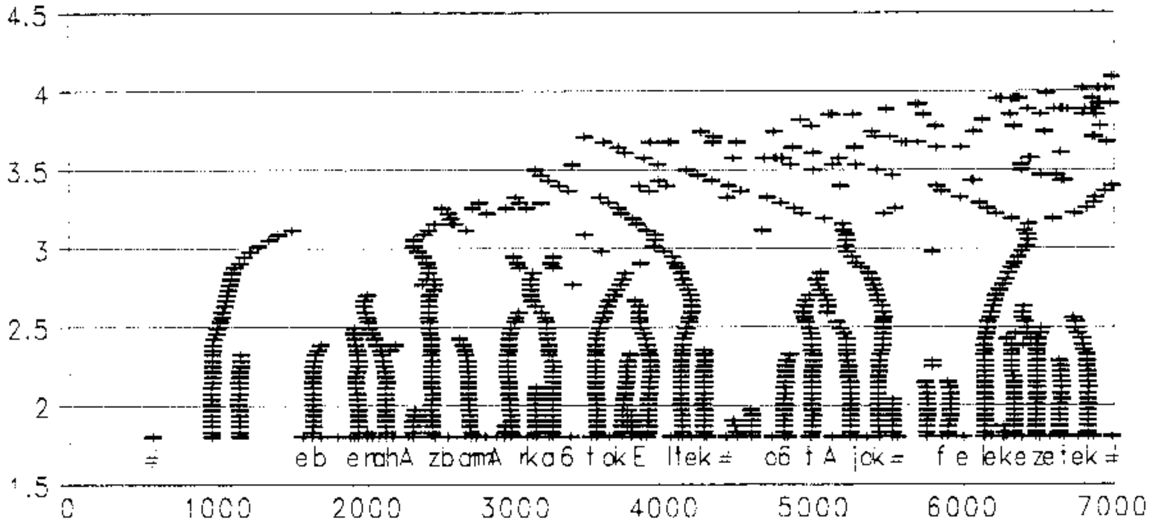
Syntagm 13

```
eb - ben  a  ház - ban  már    kaszt - ok  él - t - ek,
eb -  en  a  hAz - ban  mAr    ka6t  - ok  El - t - ek#
this-in  the house-in  already caste-PLUR live-PRET-3pl,
   in this house       already   castes     lived,
```

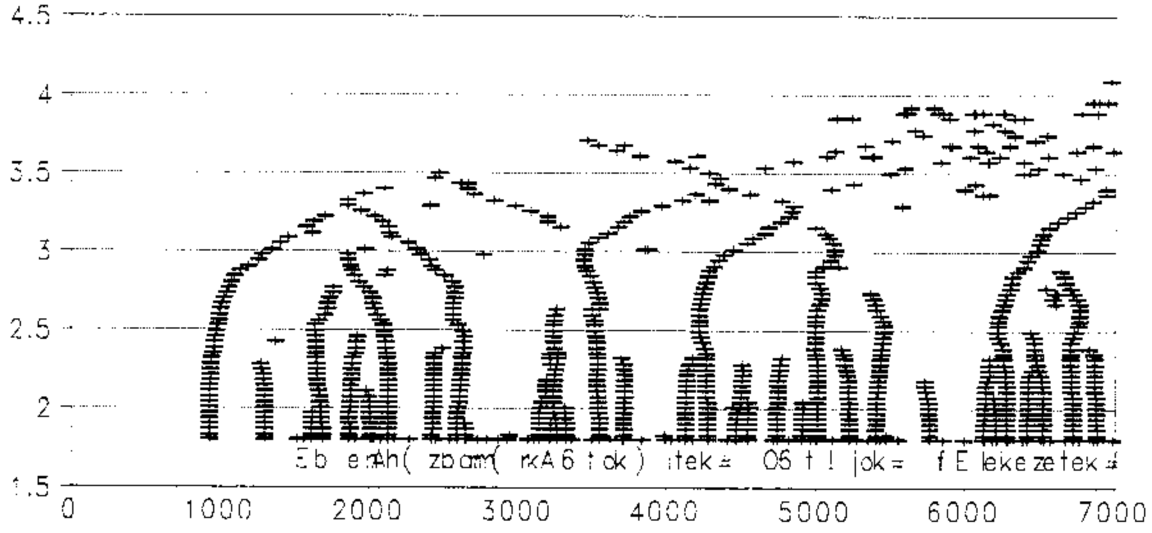
```
osztály - ok,    felekezet - ek
o6tAj   - ok#    felekezet - ek
class  - PLUR,    sect   -  PLUR
   classes,          sects.
```

"In this house there lived already castes, classes, sects".

Corpus A:



Corpus D:



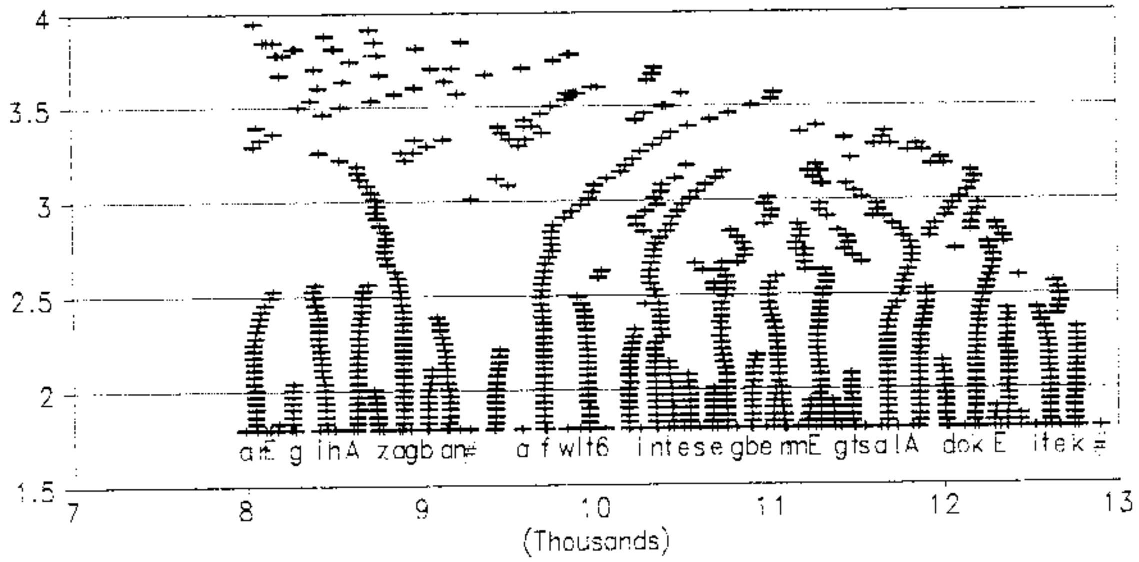
Syntagm 14

```
a   rég-i       ház - ak - ban, a föld - szint -  es -  ek - ben,
a   rEg-i       hAz - ag - ban# a fwlt -  6int -  es -  eg - ben,
the old-ADJ     house-PLUR-in, the ground-surface-ADJ- PLUR- in,
the  old        houses - in,   the single-story - houses - in,
```

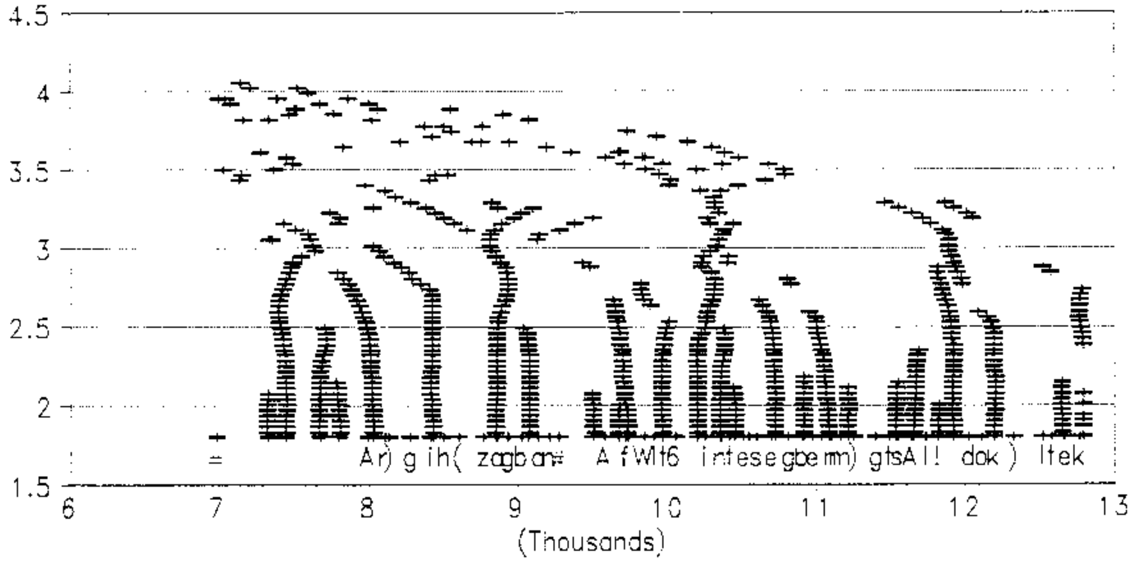
```
még  család - ok  él - t - ek
mEg  tsalAd - ok  El - t - ek
still family-PLUR live-PRET-3pl
still  families     lived
```

"In the old houses, in the single-story houses, there lived still families",

Corpus A:



Corpus D:



Syntagm 15

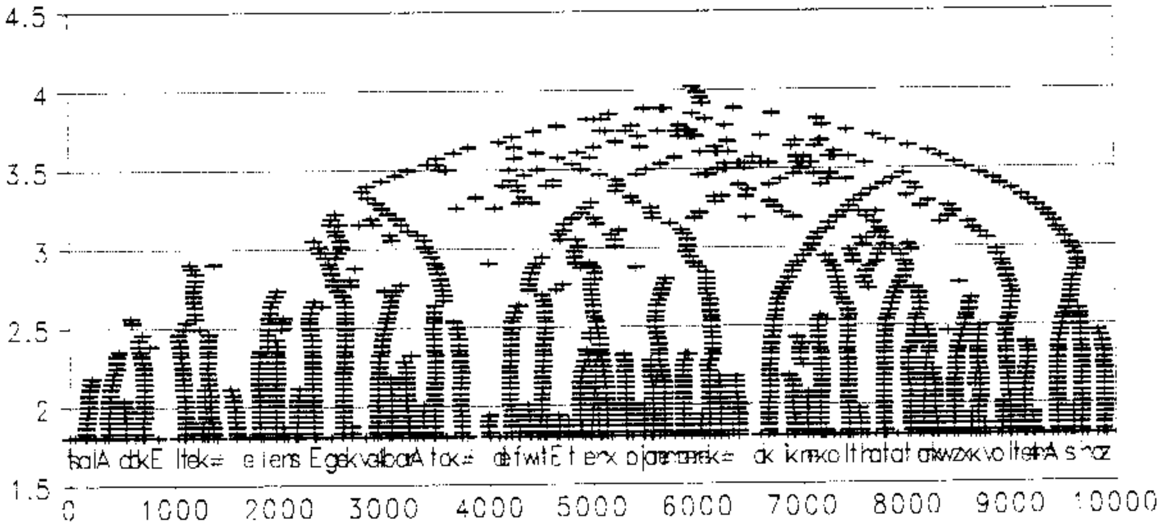
ellen - ség - ek vagy barát - ok, de föltét - len - ül olyan
elen - sEg - ek va4 barAt - ok# de fwltEt - len - xl ojan
against-NOM-PLUR or friend-PLUR, but condition-NEG- ADV such
enemies or friends but in any case such

ember-ek, aki- k - nek old - hat-atlan köz - ük
ember-ek# aki- k - nek olt - hat-atlen kwz - xk
man-PLUR, who-PLUR-DAT loosen- POT-NEG distance-3pl
people, who unseparably distance-their

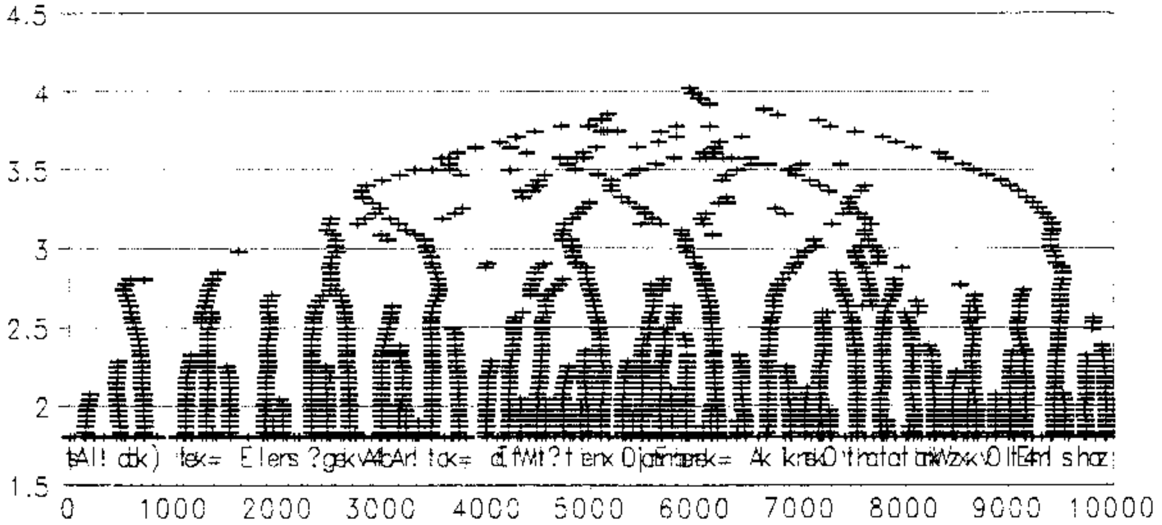
vol-t egy - más - hoz
vol-t e4 - mAs - hoz
be-PRET one-other - to
were to each other

"enemies or friends, but in any case such people who were unseparably tied to each other".

Corpus A:



Corpus D:



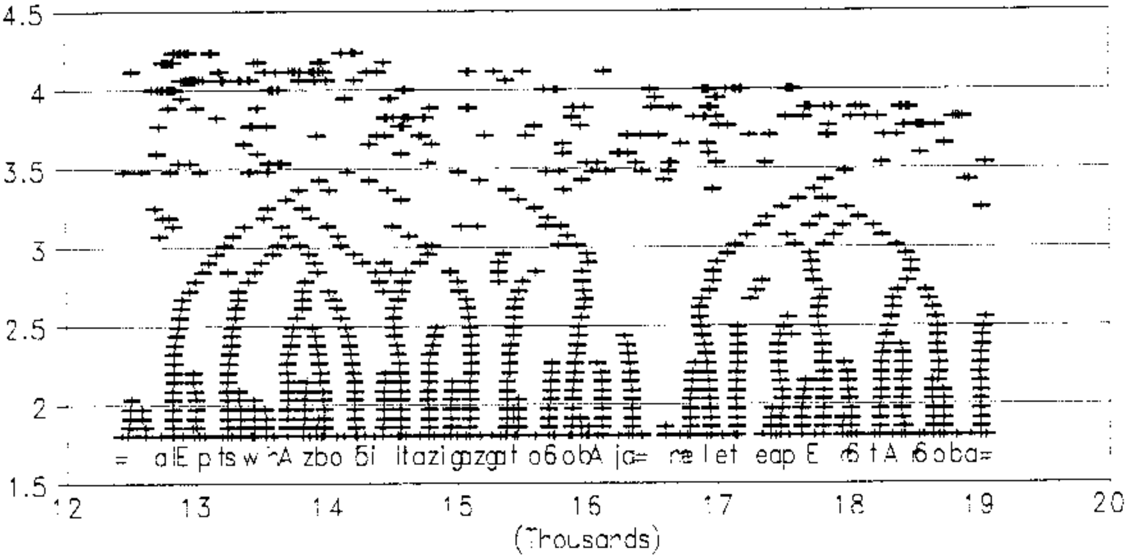
Syntagm 16

a	lép -csõ -ház - ból	nyíl - t	az	igaz-gat- ó
a	lEp -tsw -hAz - bol	5il - t	az	igaz-gat- o
the	step-tube-house-from	open-PRET	the	true-VB -PR.PT
the	staircase - from	opened	the	director

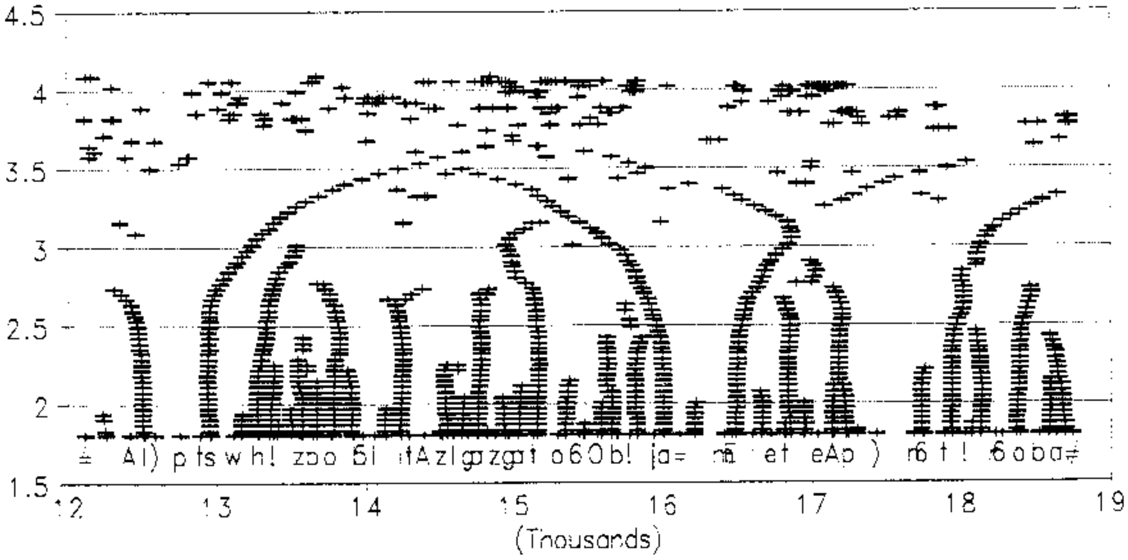
szobá-ja,	mell -ett-e	a	pénz -tár - szoba
6obA -ja,	mel - et -e	a	pEn6 -tAr - 6oba
room-3sg,	chest-PP-3sg	the	money-store-room
room-his,	beside-it		the cashier's room

"from the staircase one could enter the director's room, beside it was the cashier's room"

Corpus A:



Corpus D:



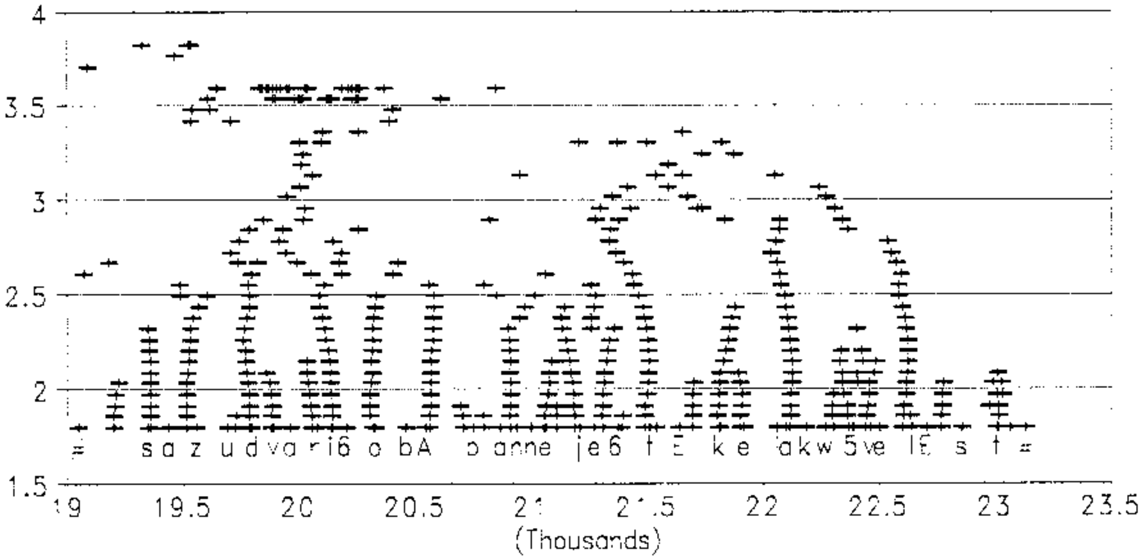
Syntagm 17

s	az	udvar - i	szobá-ban	hely - ez - t - ék	el
s	az	udvar - i	6obA -ban	hej - e6 - t - Ek	el
and	the	backyard-ADJ	room - in	place- VB -PRET-3pl	PERF
and	the	backyard	room - in	placed - they	had

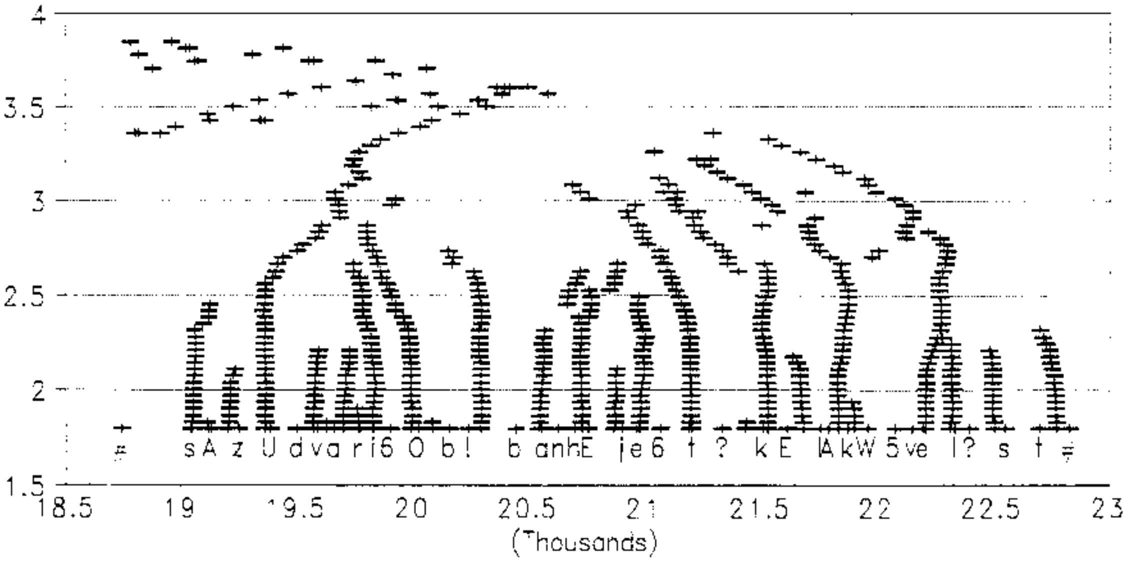
a	könyv- el - és - t
a	kw5v - el - Es - t
the	book- VB -NOM -ACC
the	accountancy

"and in the room facing the backyard they had placed the accountancy".

Corpus A:



Corpus D:



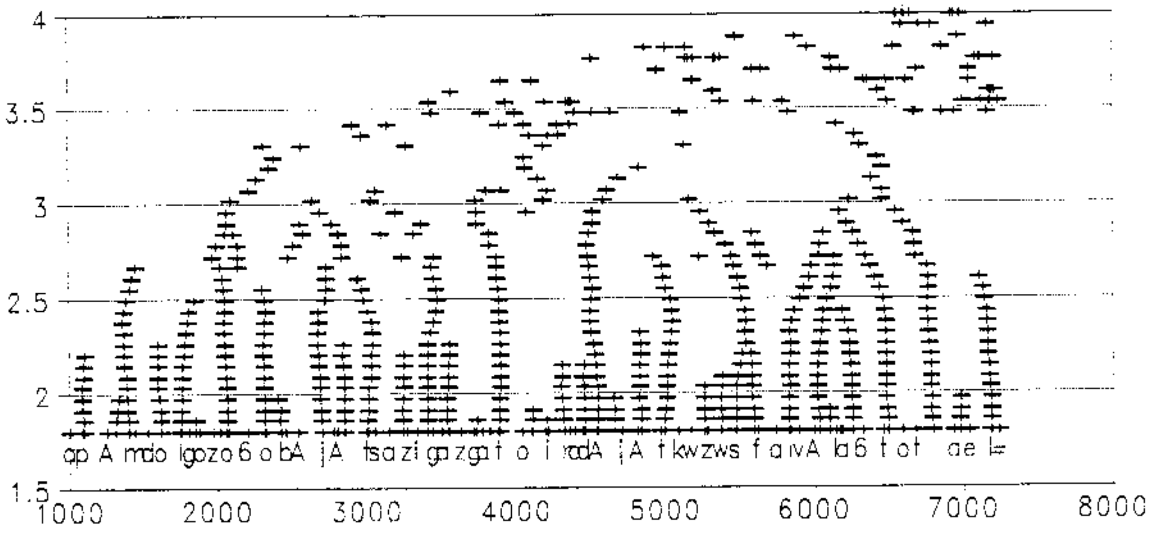
Syntagm 18

```
apá   - m  dolg    - oz   - ó - szobá - já - t  s  az  igaz -
apA   - m  dolg    - oz   - o - 6obA  - jA - t  s  az  igaz -
father-1sg matter- VB -PR.PT.-room -3sg-ACC and the true -
my father's working  -  room  (ACC)         and  the
```

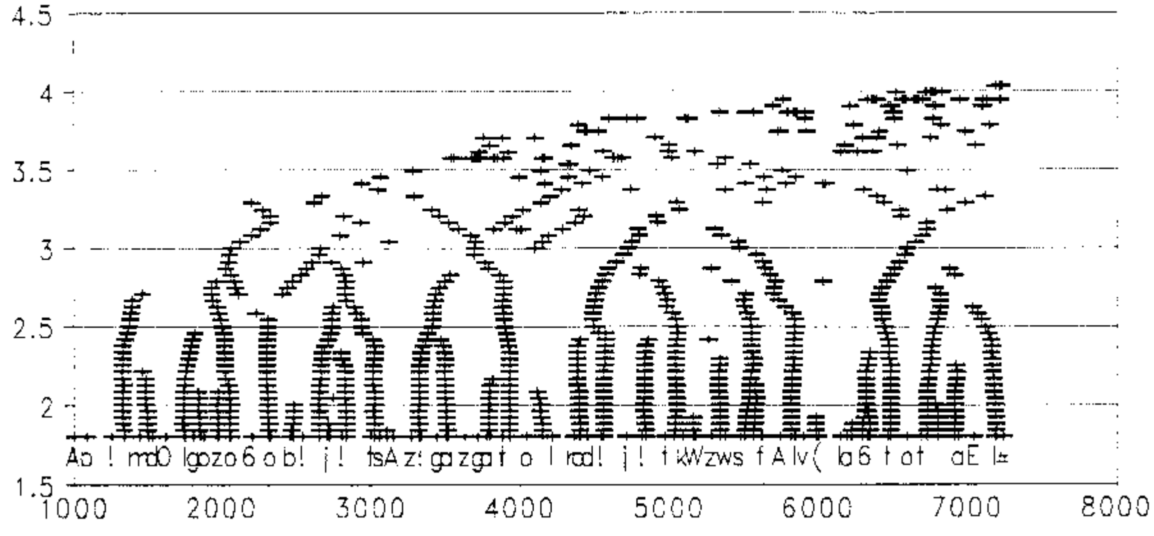
```
gat- ó    irodá - já - t  közös  fal  válasz - t - ott - a  el
gat- o    irodA - jA - t  kwzws  fal  vAla6  - t - ot  - a  el
VB -PR.PT office-3sg-ACC  common wall answer- VB -PRET-3sg PERF
director's office (ACC)   common wall   separated
```

"A common wall separated my father's study and the director's office",

Corpus A:



Corpus D:

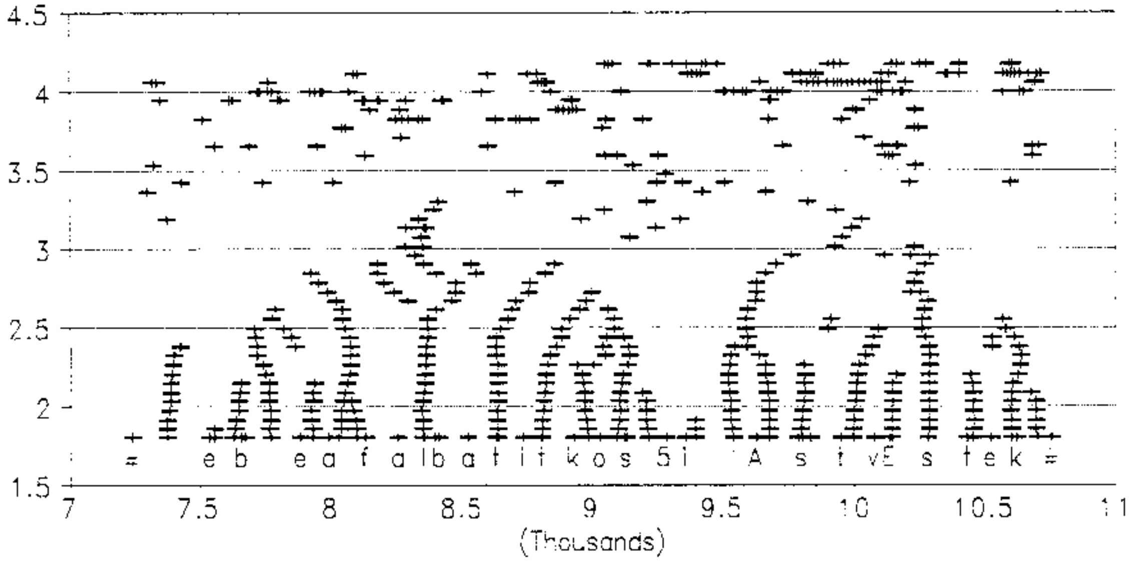


Syntagm 19

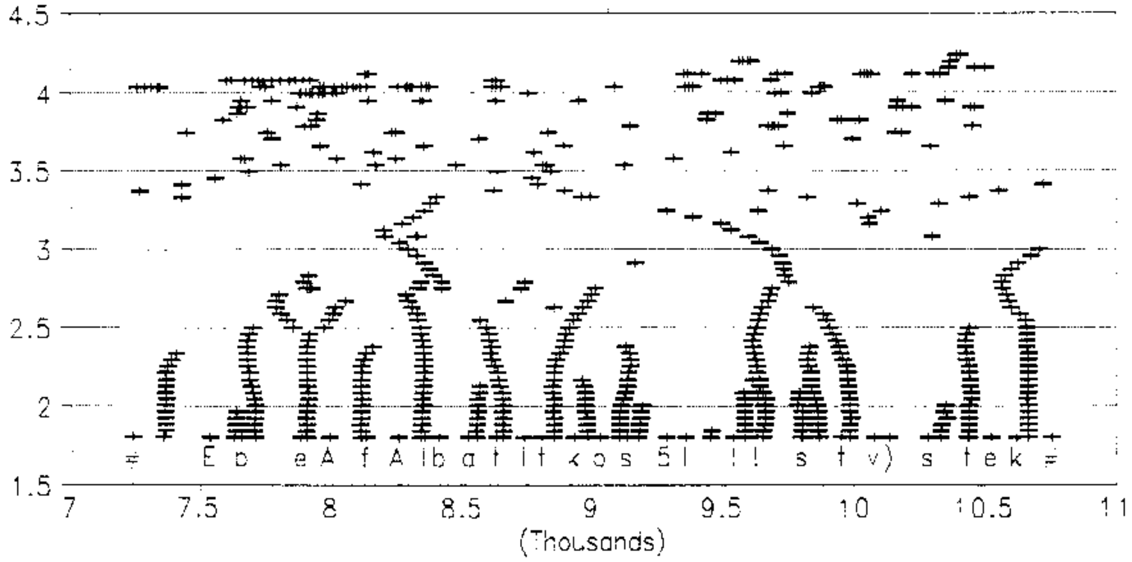
eb - be	a	fal -ba	titk - os	nyíl-ás - t	vés - t - ek
eb - e	a	fal -ba	titk - os	5il -As - t	vEs - t - ek
this-into	the	wall-into	secret-ADJ	open-NOM-ACC	carve-PRET-3pl
into is		wall	secret	opening	carved-they

"in this wall they made a secret opening",

Corpus A:



Corpus D:



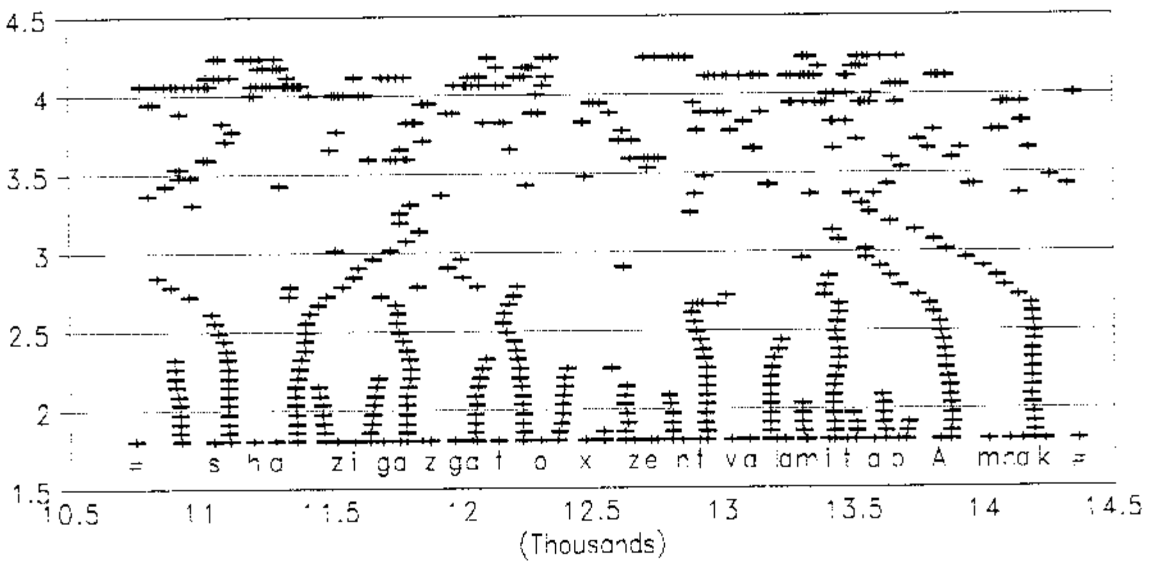
Syntagm 20

s	ha	az	igaz-gat- ó	üz - en - t	vala - mi - t
s	ha	z	igaz-gat- o	xz - en - t	vala - mi - t
and	if	the	true-VB-PR.PT	drive-VB-PRET	some - what-ACC
and	if	the	director	sent	something

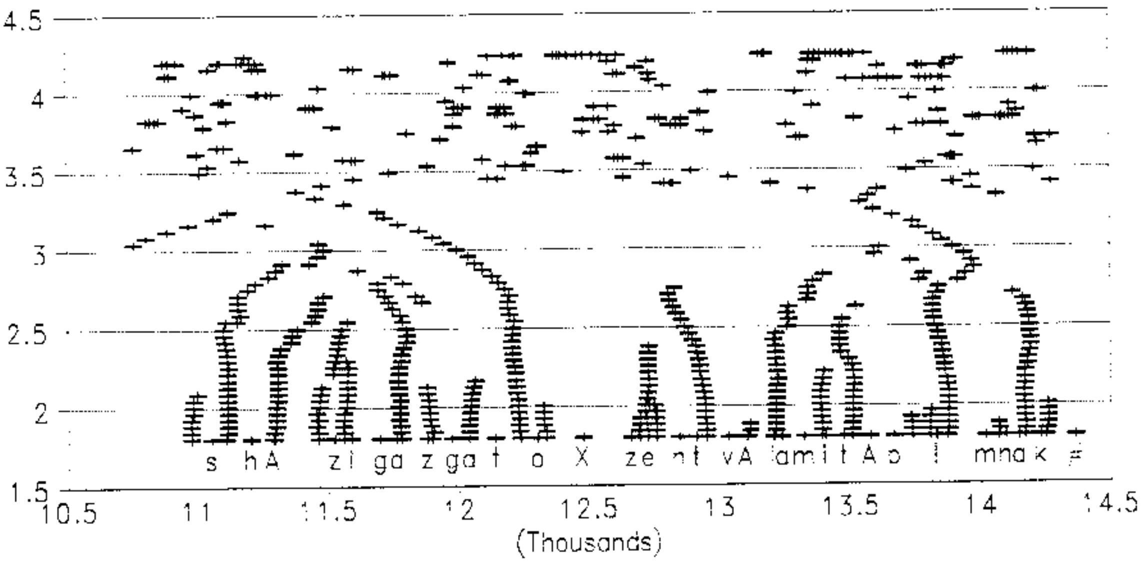
apá - m - nak
apA - m - nak
father-1sg- DAT
to my father

"and if the director should send something to my father",

Corpus A:



Corpus D:



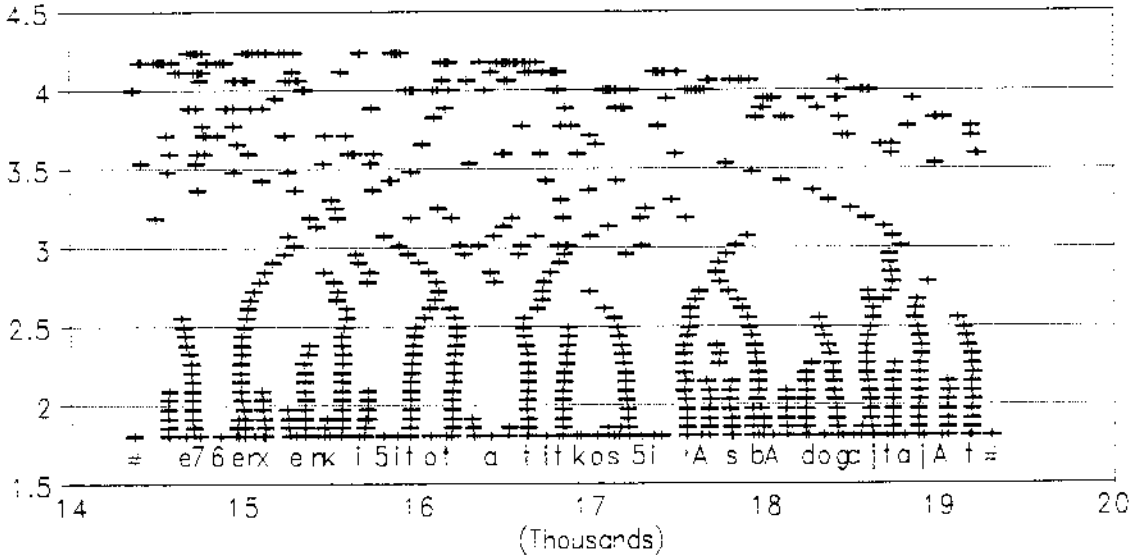
Syntagm 21

egy-szer-û - en ki-nyit-ott - a   a titk - os
e7 -6er -x - en ki-5it -ot - a (a) titk - os
one-ADV-ADJ-ADV out-open-PRET-3sg the secret-ADJ
simply   up-opened-he   the   secret

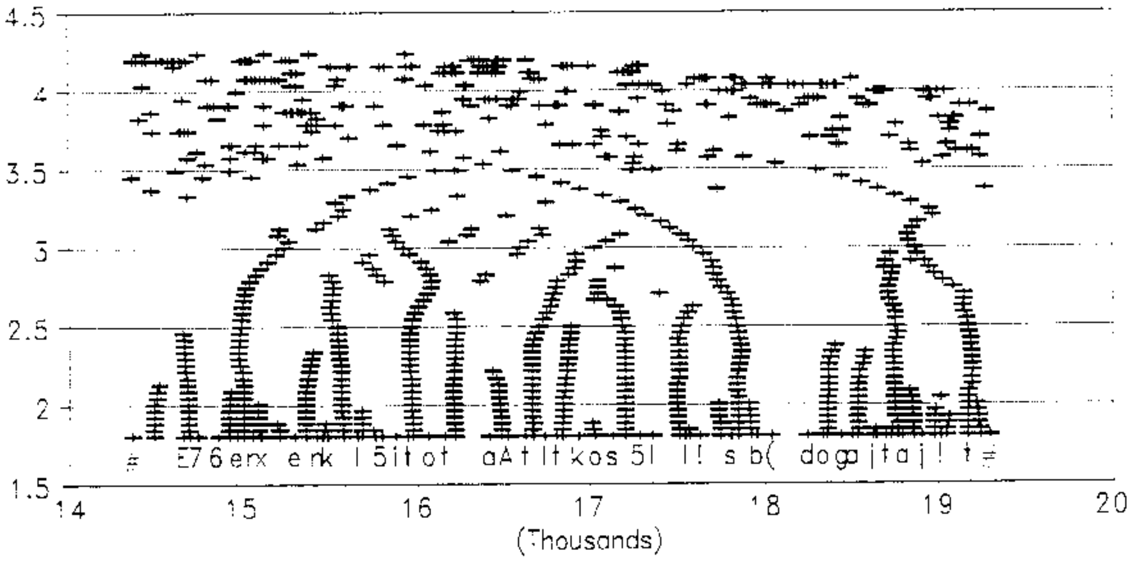
nyíl- ás   bádog   - ajta-já - t
5il - As   bAdog   - ajta-jA - t
open-NOM sheet-iron- door-3sg-ACC
opening's   sheet-iron-door

"he simply opened up the secret opening's sheet iron door",

Corpus A:



Corpus D:



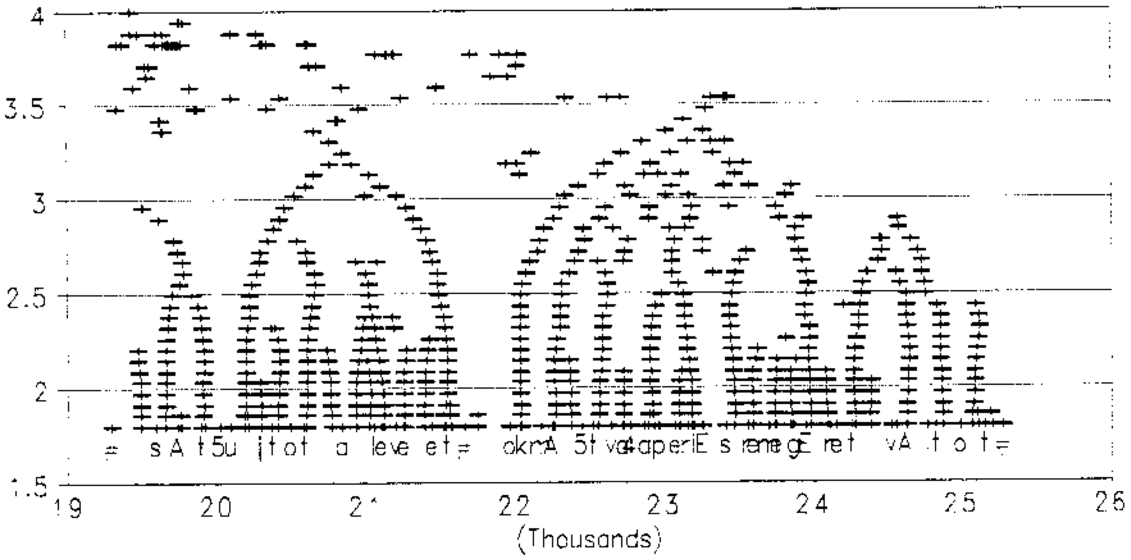
Syntagm 22

s	ât -nyújt-ott - a	a level - et,	ok - mány- t	vagy
s	At -5ujt -ot - a	(a) level - et#	ok - mA5 - t	va4
and	over-pass-PRET-3sg	the letter-ACC,	reason-NOM -ACC	or
and	handed-over-he	the letter,	document	or

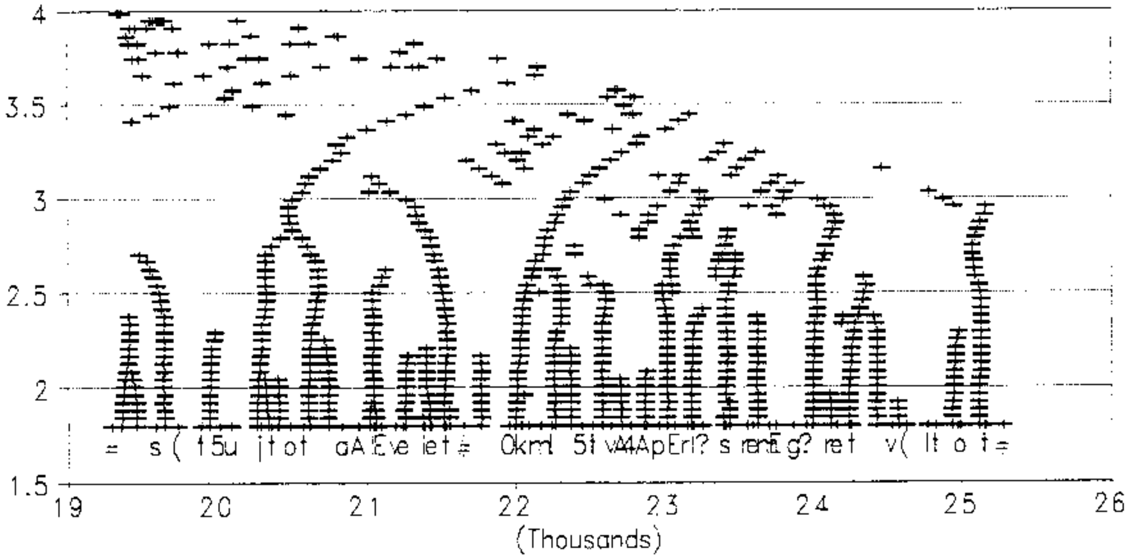
a	per - l -és -re	meg - ér - ett	vált - ó - t
a	per - l -Es -re	meg - Er - et	vAlt - o - t
the	process-VB-NOM-to	PERF-reach-P.PT	change-NOM-ACC
the	for a process	ripened	bill of exchange

"and he handed over the letter, the document or the bill of exchange ready for a process".

Corpus A:



Corpus D:



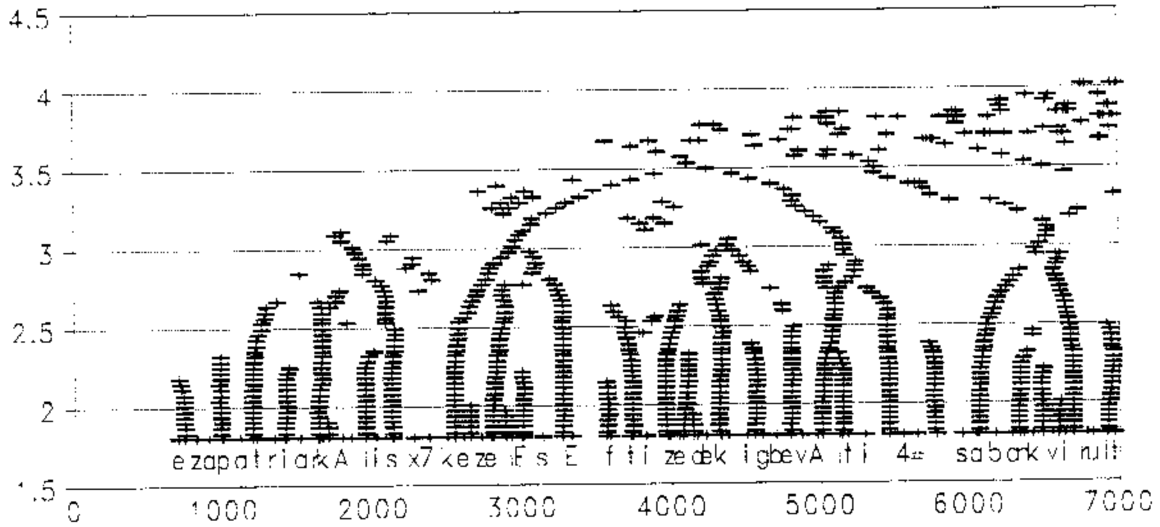
Syntagm 23

ez	a	patriarch-ál -is	ügy - kez-el- és	év -tiz-
ez	a	patriark -Al -is	x7 - kez-el- Es	Ef -tiz-
this	the	patriarch-VB-ADJ	matter-hand-VB-NOM	year-ten-
this		patriarchal	management	decades-

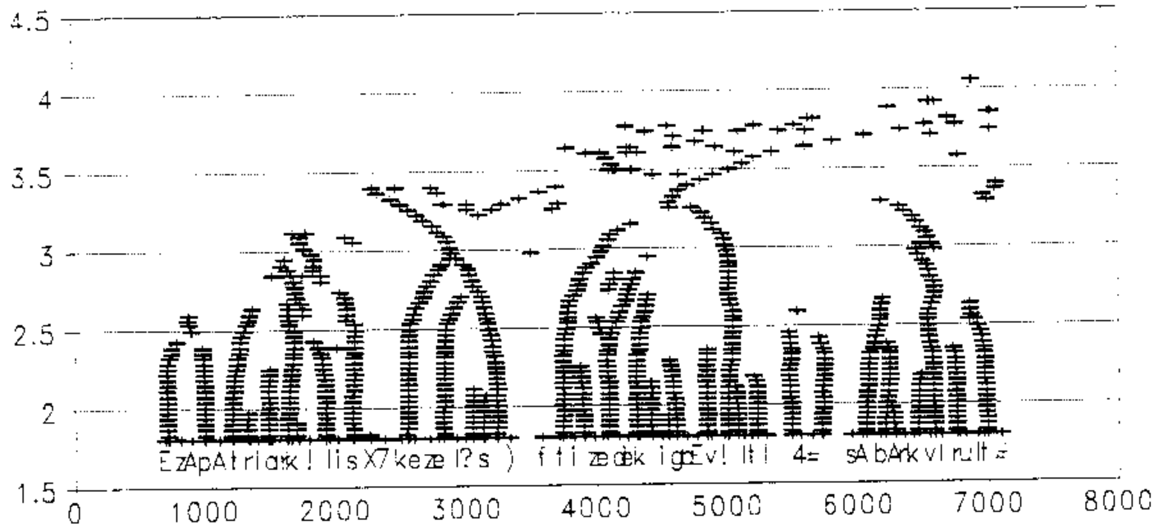
ed - ek - ig	be - vál - t	így,	s	a	bank	virul - t
ed - ek - ig	be - vAl - t	i4 #	s	a	ba9k	virul - t
NOM-PLUR-untill	in-become-PRET	so,	and	the	bank	blossom-PRET
for	worked well	so,	and	the	bank	flourished

"This patriarchal management worked well in this way for decades, and the bank flourished".

Corpus A:



Corpus D:

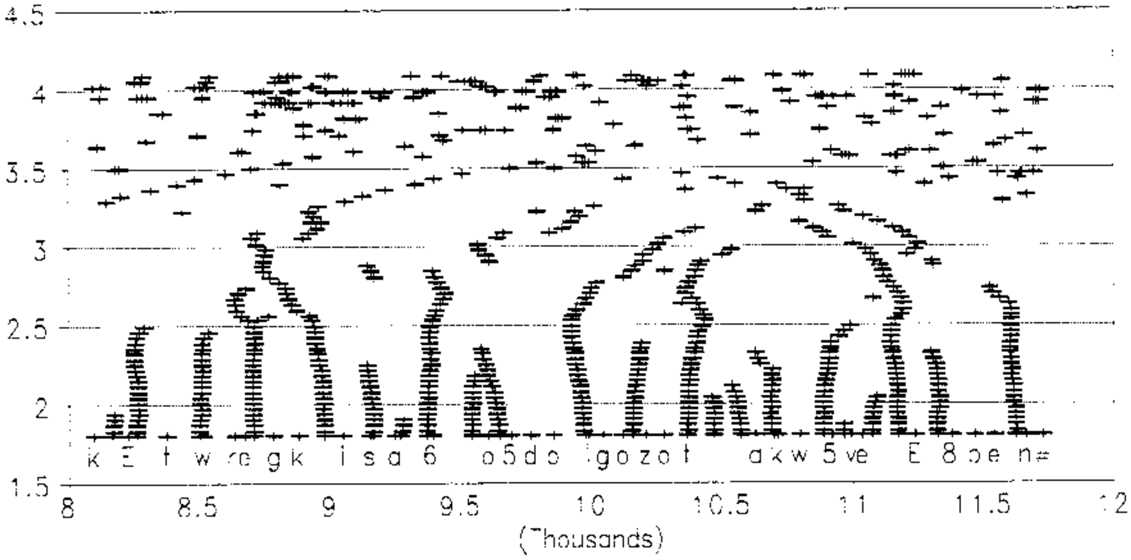


Syntagm 24

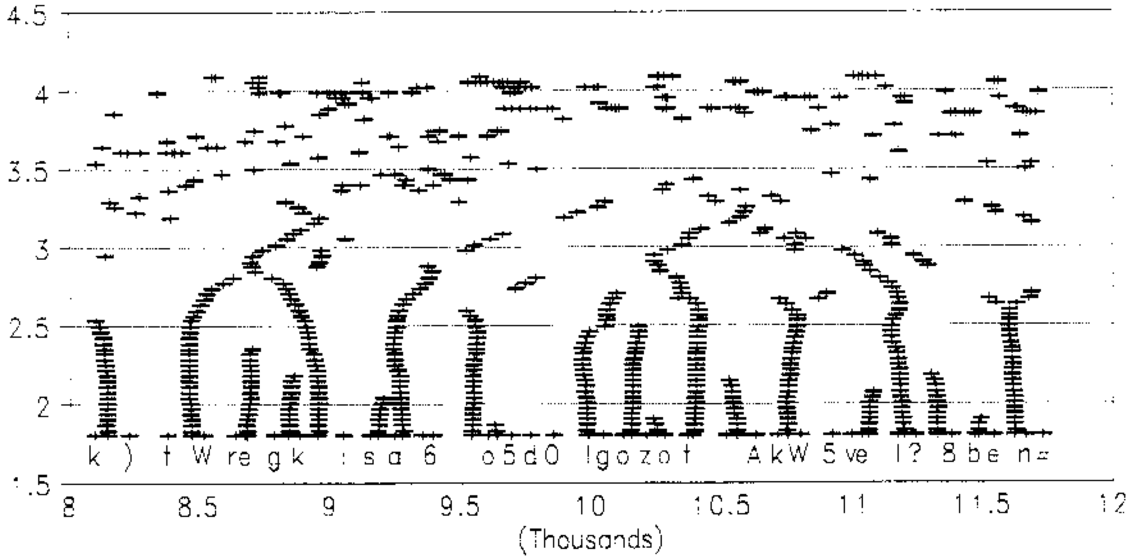
két	öreg	kis -asszony	dolg -oz- ott	a	könyv-el-és -ben
kEt	wreg	kis -a6o5	dolg -oz- ot	a	kw5v -el-E8 -ben
two	old	small-woman	matter-VB-PRET	the	book -VB-NOM-in
two	old	women	worked	the	accountancy -in

"Two old women worked in the accountancy",

Corpus A:



Corpus D:



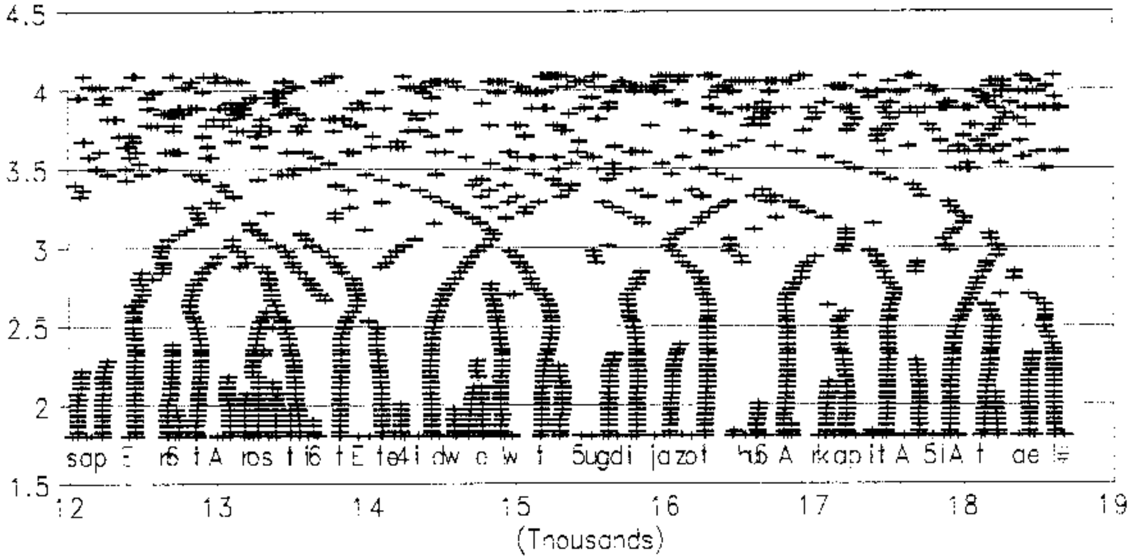
Syntagm 25

s	a	pénz - tár - os	tiszt - é - t	egy	idô	elôtt
s	a	pEn6 - tAr - os	ti6t - E - t	e4	idw	elwt
and	the	money-store-ADJ	office-3sg-ACC	a	time	before
and	the	cashier's	office	a	before-time	

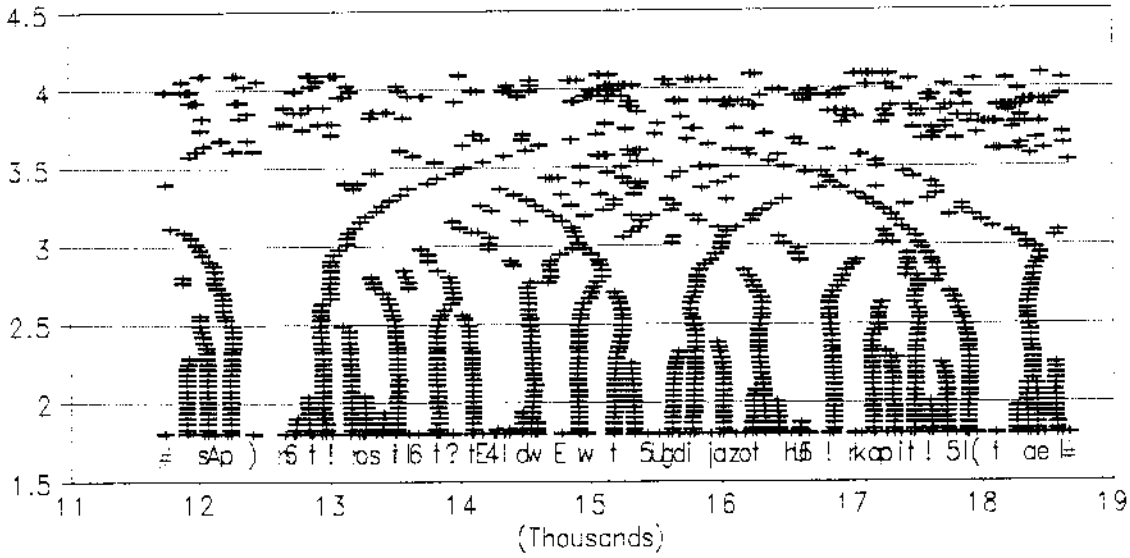
nyug - díj-az-ott	huszár	kapitány	lát-ta	el
5ug - dij-az-ot	hu6Ar	kapitA5	lAt- a	el
rest-prize-VB-PRT	cavalry	captain	see-PRET	PERF
retired	cavalry	captain	administered	

"and a cavalry captain, who had retired before time, administered the cashier's office".

Corpus A:



Corpus D:



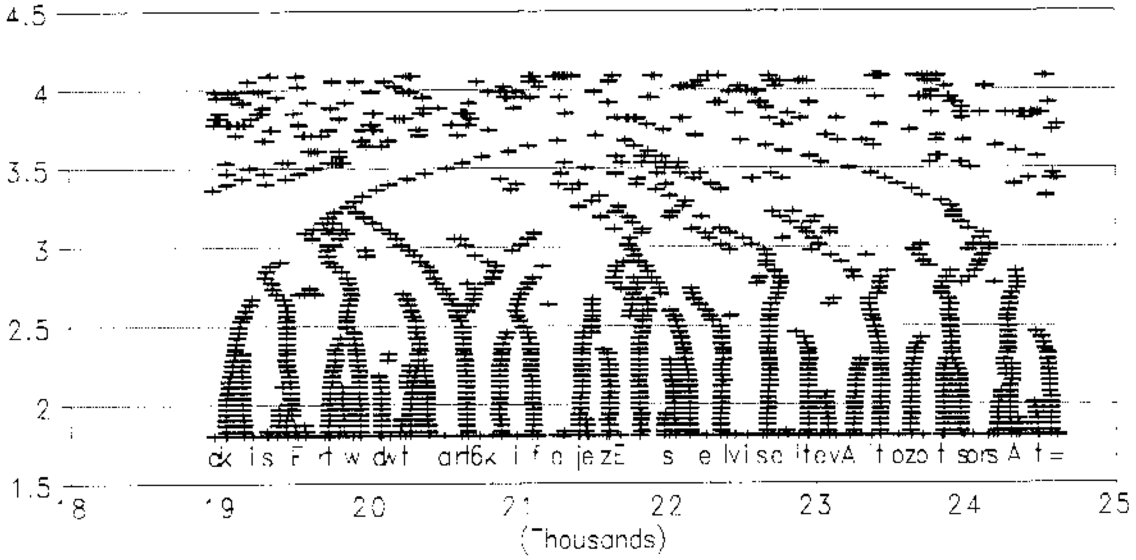
Syntagm 26

```
aki  sért - ő   - d-ött  arc- ki-fej -ez-és - sel
aki  sErt - w   - d-wt   art6-ki-fej -ez-Es -  el
who offend-PR.PT-VB-PRT face-out-head-VB-NOM-INSTR
who     pained              expression  -     with
```

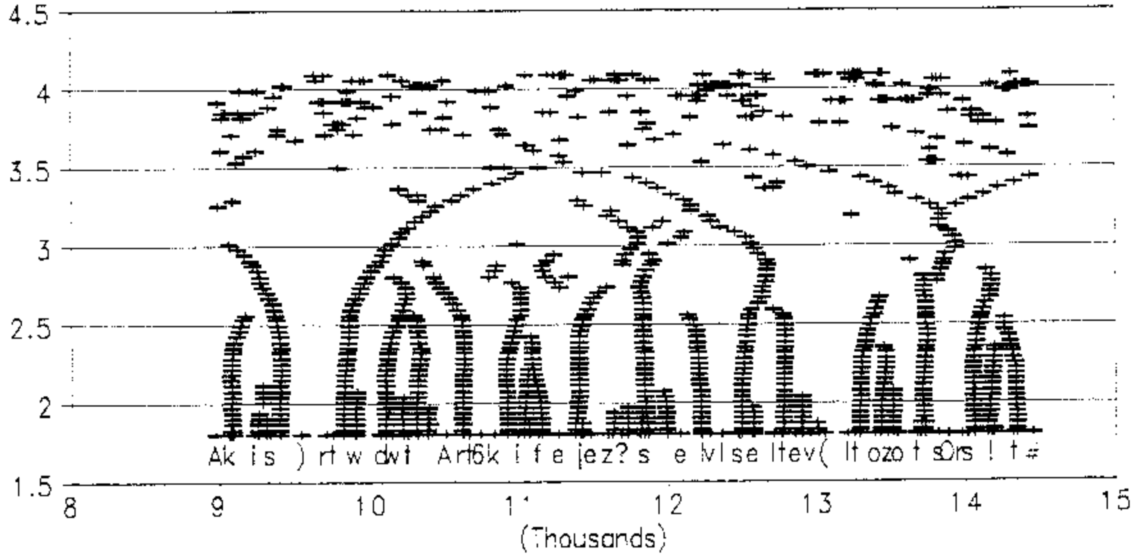
```
visel - te  vált  -oz-ott  sors- á - t
visel - te  vAlt  -oz- ot  sors- A - t
endure-PRET change-VB-PRT  fate-3sg-ACC
   endured  changed        fate-his
```

"who with a pained expression endured his new fate",

Corpus A:



Corpus D:



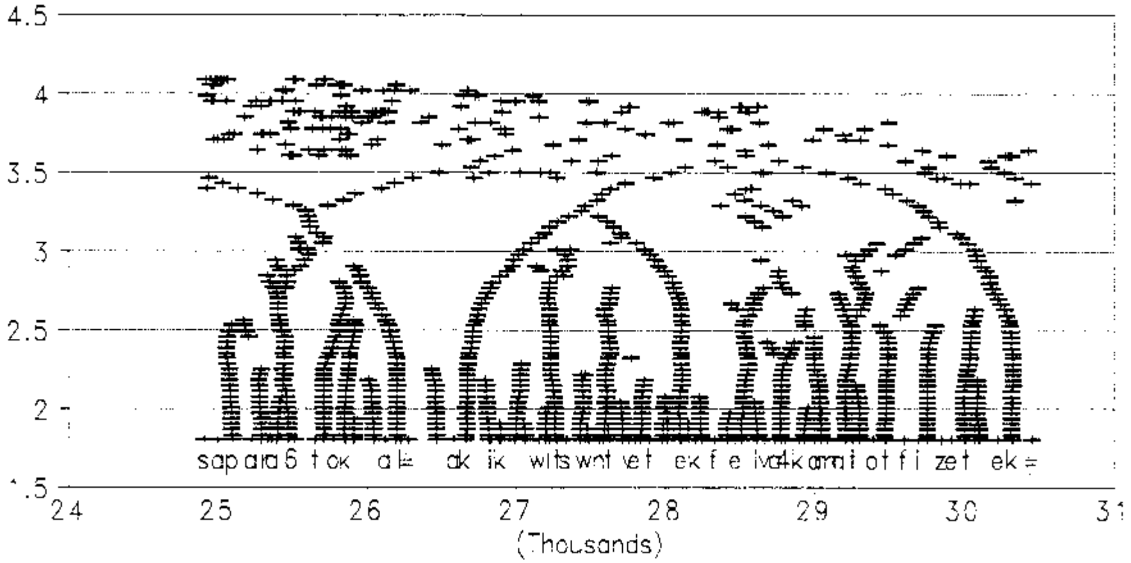
Syntagm 27

s	a	paraszt- ok - kal,	aki-k	kölcsön-t	vet - t - ek
s	a	para6t - ok - al#	aki-k	(k)wltswn-t	vet - ek
and	the	peasant-PLUR-INSTR,	who-PLUR	loan -ACC	throw-PRET-3pl
and	the	peasants-with,	who	loan	borrowed

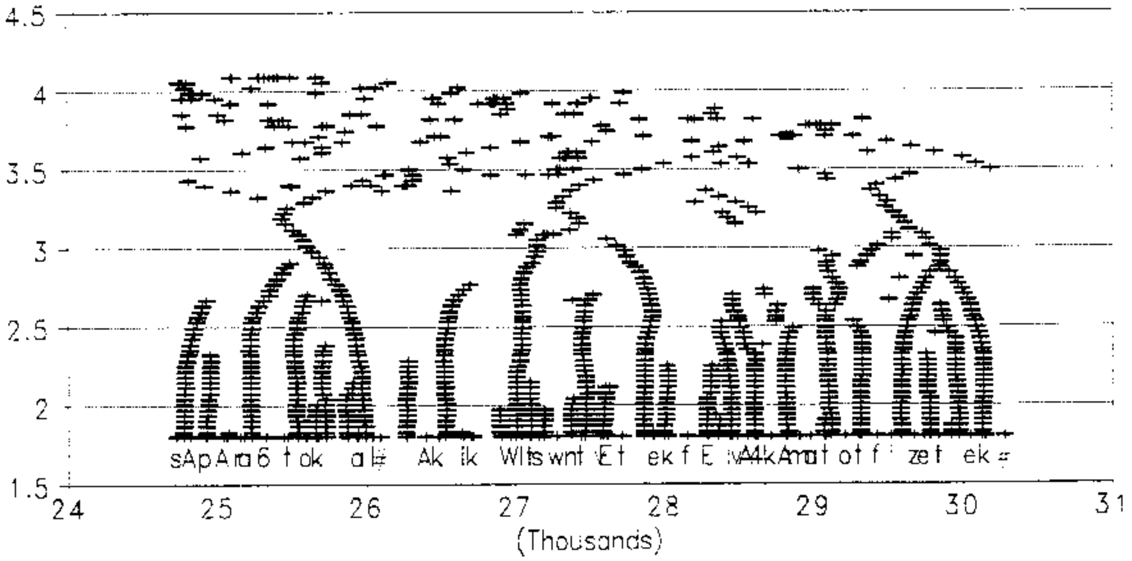
fel	vagy	kamat - ot	fizet- t - ek
fel	va4	kamat - ot	fizet - ek
up	or	interest-ACC	pay -PRET-3pl
	or	interest	paid

"and to the peasants, who borrowed money or paid interest",

Corpus A:



Corpus D:

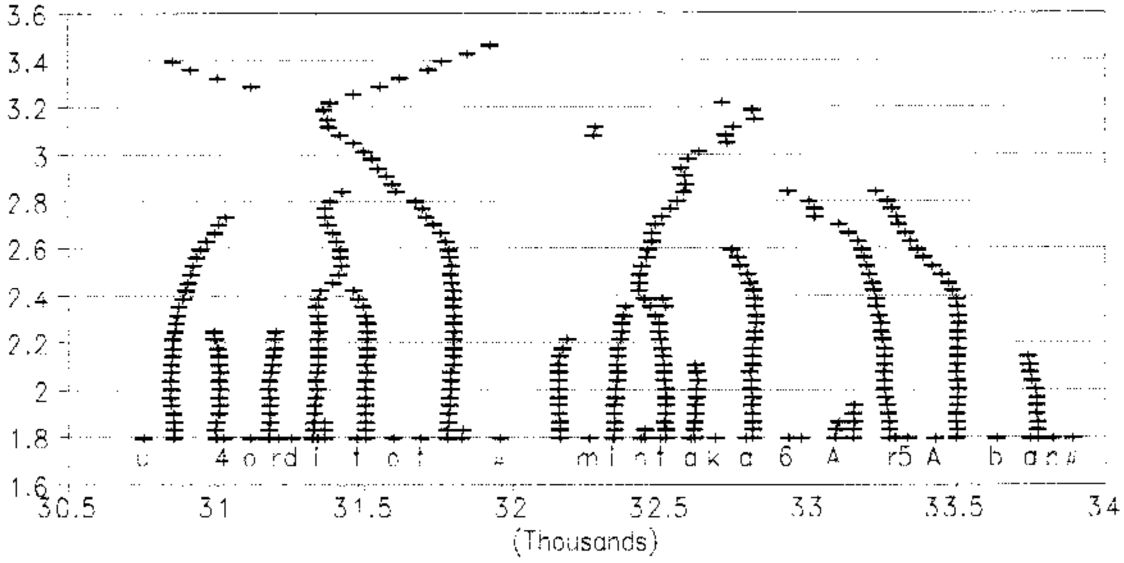


Syntagm 28

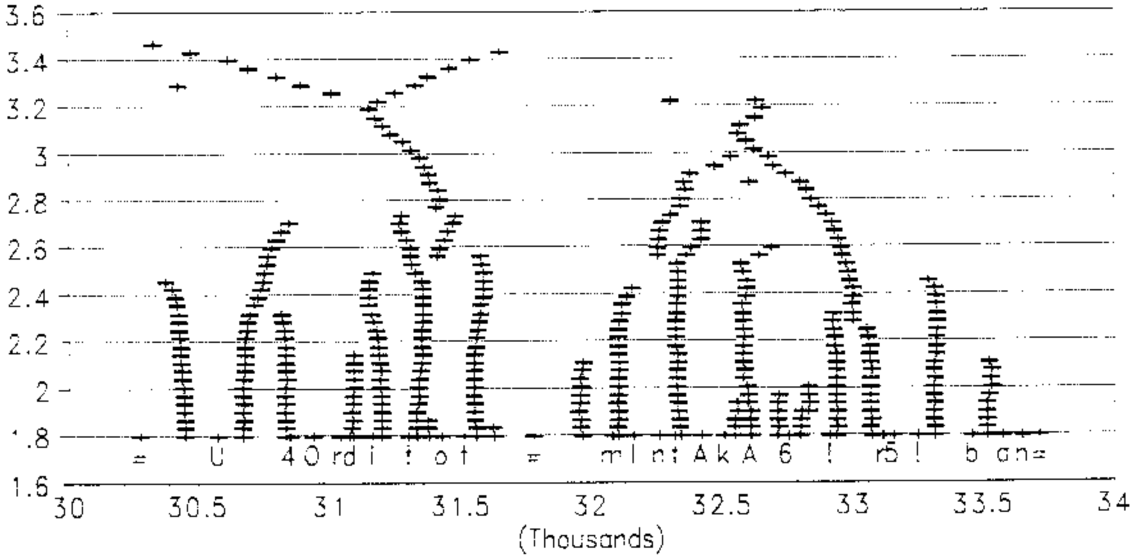
úgy	ordít-ott,	mint	a	kaszárnyá-ban
u4	ordit-ot #	mint	a	ka6Ar5A - ban
so	shout-PRET,	like	the	barrack - in
so	shouted,	like	the	barrack - in

"he shouted as in the barrack".

Corpus A:



Corpus D:



APPENDIX B

This contains the rules for redefinition from text to speech approximation in corpora A and C, as well as the duration values for the corpus symbols. The redefinition for corpus C has been applied to the readily defined corpus A, which means that the redefinition for corpus C has applied over all boundaries which are not separated by a #-symbol.

The program is designed to choose the largest context which matches the input, and it automatically converts normal majuscules into minuscules before the redefinition.

For the duration values, the sound to be computed is given in brackets. The numbers indicate duration in milliseconds.

For technical reasons, some of the durations have been split. For example, a short [o] with no context is given as 107 milliseconds (see the eleventh entry in the first column). Immediately below there is the entry [o]o=63. Thus, if in the corpus a character 'o' is followed by another 'o', it has duration 63. If the following has no matching context, this will have duration 107, which means that they will have a duration of 170 ms altogether. A long [o:] (which thus has duration 170 ms as opposed to the short [o] with duration 107 ms) has been coded in this way in order for the symbol 'o' to be counted in both cases. For the long vowels 'A' and 'E', which are defined as acoustically different from their short counterparts, this has not been necessary.

For technical reasons similar to the vowels, the affricates have been coded as series of two or three symbols. The duration of an affricate is thus the sum of its parts.

The program chooses the largest matching context. If, as may be the case for the affricates, there are both a lefthand and a righthand context, the size of the context is the sum of the lefthand and the righthand context. If this sum is equally large as another matching context, the program will choose the largest righthand context.

Redefinition for corpus A

```
á=A
Á=A
é=E
É=E
í=ii
ì=ii
ó=oo
ò=oo
ö=w
Ö=w
ȏ=ww
º=ww
ú=uu
ù=uu
ü=x
Ü=x
û=xx
[illegible]=xx
cs=ts
dzs=d8
gy=4
ny=5
sz=6
ty=7
zs=8
ccs=tts
ddzs=dd8
ggy=44
nny=55
ssz=66
tty=77
zzs=88
nk=9k
ng=9g
ly=j
lly=jj
x=k6
w=v
stb=satwbbi
 kg=kilo
 g.=gramm
pl.=pEldAul
kb.=kwrxlbelxl
 ill.=illetve
 dk.=deka
dkg=deka
 l.=liter
 km=kilomEter
cm=centimEter
usd=ue6dollAr
%=6AzalEk
 u.=utca
gfk=gEefkA

bp=pp
bf=pf
bt=pt
bsz=p6
bc=pt6
bcs=pts
bs=ps
bsz=p6
bty=p7
bk=pk
bh=ph
vp=fp
vt=ft
vsz=f6
vc=ft6
vcs=fts
vs=fs
vty=f7
vk=fk
dp=tp
df=tf
dt=tt
dsz=t6
dc=tt6
ds=ts
dcs=tts
dty=t7
dk=tk
dh=th
zp=6p
zt=6t
zsz=66
zc=6t6
zcs=6ts
zty=67
zk=6k
zh=6h
dzp=t6p
dzf=t6f
dzt=t6t
dzsz=t66
dzc=tt6
dzs=tts
dzcs=t6ts
dzty=t67
dzk=t6k
dzh=t6h
zsp=sp
zsf=sf
zst=st
zssz=s6
zsc=st6
zss=ss

dzsf=tsf
dzst=tst
dzssz=t66
dzsc=tts
dzss=tts
dzsty=ts7
dzsk=tsk
dzsh=tsh
gyp=7p
gyf=7f
gyt=7t
gysz=76
gyc=7t6
gys=7s
gycs=7t6
gyty=77
gyk=7k
gyh=7h
gp=kp
gf=kf
gt=kt
gsz=k6
gc=kt6
gs=ks
gcs=kts
gty=k7
gk=kk
gh=kh
pb=bb
pd=bd
pz=bz
pdz=bdz
pzs=b8
pdzs=bd8
pgy=b4
pg=bg
fz=vz
fzs=v8
tb=db
td=dd
tz=dz
tdz=ddz
tzs=d8
tdzs=dd8
tgy=d4
tg=dg
szb=zb
szd=zd
szz=zz
szdz=zdz
szzs=z8
cb=dzb
cd=dzd

cg=dzg
sb=8b
sd=8d
sdz=8dz
sdzs=8d8
sgy=84
sg=8g
csb=d8b
csd=d8d
csdz=d8dz
csgy=d84
csg=d8g
tyb=4b
tyd=4d
tyz=4z
tydz=ddz
tyzs=48
tydzs=4d8
tygy=44
tyg=4g
kb=gb
kd=gd
kz=gz
kdz=gdz
kzs=g8
kdzs=gd8
kgy=g4
kg=gg
lj=jj
nyj=55
tyj=77
szs=ss
nb=mb
np=mp
nm=mm
ngy=54
nty=57
.=##########
?=##########
!=##########
:=######
;=#######
,=###
,,=
 =
.,=
 - =#####
-=
+=plu66
&=Es
.-=
```

kft=kAeftE
rt=ertE
db=darab

zscs=sts
zsty=s7
zsk=sk
zsh=sh
dzsp=tsp

cz=ddz
cdz=ddz
czs=dd8
cdzs=dzd8
cgy=dz4

Additional redefinition for corpus C

nk=9k
ng=9g
np=mp
nb=mb
n4=54
n7=57
bp=pp
bf=pf
bt=pt
b6=p6
bs=ps
b6=p6
b7=p7
bk=pk
bh=ph
vp=fp
vt=ft
v6=f6
vs=fs
v7=f7
vk=fk
dp=tp
df=tf
dt=tt
d6=t6
ds=ts
d7=t7

dk=tk
dh=th
zp=6p
zt=6t
z6=66
z7=67
zk=6k
zh=6h
dzp=t6p
dzf=t6f
dzt=t6t
dz6=t66
dzs=tts
dz7=t67
dzk=t6k
dzh=t6h
8p=sp
8f=sf
8t=st
86=s6
8s=ss
87=s7
8k=sk
8h=sh
4p=7p
4f=7f
4t=7t

46=76
4s=7s
47=77
4k=7k
4h=7h
gp=kp
gf=kf
gt=kt
g6=k6
gs=ks
g7=k7
gk=kk
gh=kh
pb=bb
pd=bd
pz=bz
p8=b8
p4=b4
pg=bg
fz=vz
f8=v8
tb=db
td=dd
tz=dz
t8=d8
t4=d4
tg=dg

6b=zb
6d=zd
6z=zz
68=z8
t6b=dzb
t6d=dzd
t64=dz4
t6g=dzg
sb=8b
sd=8d
s4=84
sg=8g
tsb=d8b
tsd=d8d
ts4=d84
tsg=d8g
7b=4b
7d=4d
7z=4z
74=44
7g=4g
kb=gb
kd=gd
kz=gz
k4=g4
kg=gg
y=i

DURATION VALUES

```
[e]=109
[E]=178
[i]=106
[i]i=87
[w]=111
[w]w=102
[x]=105
[x]x=83
[a]=110
[A]=196
[o]=107
[o]o=63
[u]=109
[u]u=94
[e]p=86
[E]p=172
[i]p=74
[i]ip=157
[w]p=102
[w]wp=81
[x]p=75
[x]xp=66
[a]p=94
[A]p=180
[o]p=68
[o]op=104
[u]p=117
[u]up=62
[e]b=121
[E]b=131
[i]b=170
[i]ib=10
[w]b=125
[w]wb=47
[x]b=90
[x]xb=57
[a]b=137
[A]b=204
[o]b=129
[o]ob=37
[u]b=83
[u]ub=37
[e]t=98
[E]t=143
[i]t=68
[i]it=64
[w]t=98
[w]wt=98
[x]t=90
[x]xt=59

[w]m=94
[w]wm=102
[x]m=136
[x]xm=84
[a]m=84
[A]m=185
[o]m=90
[o]om=106
[u]m=102
[u]um=86
[e]n=121
[E]n=251
[i]n=94
[i]in=130
[w]n=98
[w]wn=76
[x]n=150
[x]xn=50
[a]n=75
[A]n=187
[o]n=99
[o]on=26
[u]n=75
[u]un=70
[e]5=91
[E]5=143
[i]5=141
[i]i5=16
[w]5=121
[w]w5=93
[x]5=91
[x]x5=52
[a]5=103
[A]5=200
[o]5=92
[o]o5=29
[u]5=172
[u]u5=28
[e]t6=83
[E]t6=196
[i]t6=75
[i]it6=121
[w]t6=149
[w]wt6=71
[x]t6=110
[x]xt6=90
[a]t6=117
[A]t6=235
[o]t6=149
[o]ot6=63

[b]bii=43
[b]bw=203
[b]bww=129
[b]bx=146
[b]bxx=129
[b]ba=150
[b]bA=135
[b]bo=88
[b]boo=86
[b]bu=129
[b]buu=129
[t]=123
[t]e=90
[t]E=124
[t]i=110
[t]ii=113
[t]w=140
[t]ww=149
[t]x=90
[t]xx=123
[t]a=98
[t]A=145
[t]o=124
[t]oo=165
[t]u=123
[t]uu=123
[t]t=144
[t]te=121
[t]tE=182
[t]ti=172
[t]tii=107
[t]tw=150
[t]tww=118
[t]tx=177
[t]txx=144
[t]ta=155
[t]tA=185
[t]to=143
[t]too=114
[t]tu=120
[t]tuu=136
[d]=89
[d]e=60
[d]E=165
[d]i=80
[d]ii=89
[d]w=75
[d]ww=90
[d]x=89
[d]xx=89

[z]zA=112
[z]zo=164
[z]zoo=82
[z]zu=43
[z]zuu=119
[s]=150
[s]e=98
[s]E=200
[s]i=196
[s]ii=150
[s]w=124
[s]ww=150
[s]x=143
[s]xx=170
[s]a=90
[s]A=150
[s]o=121
[s]oo=150
[s]u=204
[s]uu=150
[s]s=134
[s]se=143
[s]sE=141
[s]si=88
[s]sii=250
[s]sw=63
[s]sww=134
[s]sx=141
[s]sxx=114
[s]sa=192
[s]sA=164
[s]so=106
[s]soo=117
[s]su=94
[s]suu=134
[8]=92
[8]e=113
[8]E=92
[8]i=85
[8]ii=92
[8]w=92
[8]ww=92
[8]x=92
[8]xx=92
[8]a=75
[8]A=83
[8]o=71
[8]oo=92
[8]u=125
[8]uu=92

[m]mo=151
[m]moo=47
[m]mu=105
[m]muu=102
[5]=74
[5]e=70
[5]E=74
[5]i=74
[5]ii=74
[5]w=74
[5]ww=74
[5]x=74
[5]xx=74
[5]a=60
[5]A=94
[5]o=70
[5]oo=74
[5]u=74
[5]uu=74
[5]5=96
[5]5e=144
[5]5E=96
[5]5i=96
[5]5ii=114
[5]5w=96
[5]5ww=96
[5]5x=96
[5]5xx=75
[5]5a=68
[5]5A=76
[5]5o=100
[5]5oo=96
[5]5u=96
[5]5uu=96
[t]6=87
t[6]=86
[t]6E=106
t[6]E=106
[t]6i=91
t[6]i=90
[t]6ii=87
t[6]ii=86
[t]6w=75
t[6]w=75
[t]6ww=94
t[6]ww=94
[t]6a=75
t[6]a=74
[t]6A=87
t[6]A=87
```

```
[a]t=105
[A]t=157
[o]t=90
[o]ot=68
[u]t=141
[u]ut=10
[e]d=121
[E]d=188
[i]d=133
[i]id=91
[w]d=141
[w]wd=33
[x]d=125
[x]xd=79
[a]d=117
[A]d=205
[o]d=61
[o]od=122
[u]d=110
[u]ud=149
[e]k=98
[E]k=174
[i]k=66
[i]ik=102
[w]k=84
[w]wk=59
[x]k=90
[x]xk=130
[a]k=84
[A]k=158
[o]k=78
[o]ok=65
[u]k=62
[u]uk=134
[e]g=113
[E]g=147
[i]g=105
[i]ig=137
[w]g=136
[w]wg=154
[x]g=110
[x]xg=47
[a]g=77
[o]g=105
[o]og=40
[u]g=68
[u]ug=42
[e]f=123
[E]f=250
[i]f=133
[i]if=55
[w]f=117
```

```
[u]t6=141
[u]ut6=102
[e]dz=109
[E]dz=178
[i]dz=106
[i]idz=87
[w]dz=111
[w]wdz=102
[x]dz=105
[x]xdz=83
[a]dz=110
[A]dz=196
[o]dz=107
[o]odz=40
[u]dz=109
[u]udz=94
[e]ts=90
[E]ts=178
[i]ts=77
[i]its=116
[w]ts=111
[w]wts=124
[x]ts=105
[x]xts=83
[a]ts=125
[A]ts=196
[o]ts=129
[o]ots=28
[u]ts=110
[u]uts=102
[e]d8=109
[E]d8=178
[i]d8=106
[i]id8=87
[w]d8=111
[w]wd8=102
[x]d8=117
[x]xd8=71
[a]d8=143
[A]d8=196
[o]d8=121
[o]od8=49
[u]d8=109
[u]ud8=94
[e]tt6=83
[E]tt6=196
[i]tt6=75
[i]itt6=121
[w]tt6=149
[w]wtt6=71
[x]tt6=110
[x]xtt6=90
```

```
[d]a=60
[d]A=94
[d]o=89
[d]oo=89
[d]u=86
[d]uu=89
[d]d=123
[d]de=175
[d]dE=23
[d]di=94
[d]dii=68
[d]dw=137
[d]dww=122
[d]dx=123
[d]dxx=123
[d]da=191
[d]dA=173
[d]do=123
[d]doo=123
[d]du=126
[d]duu=123
[k]=131
[k]e=105
[k]E=131
[k]i=168
[k]ii=151
[k]w=83
[k]ww=192
[k]x=113
[k]xx=131
[k]a=105
[k]A=131
[k]o=98
[k]oo=141
[k]u=149
[k]uu=131
[k]k=122
[k]ke=66
[k]kE=183
[k]ki=85
[k]kii=102
[k]kw=170
[k]kww=61
[k]kx=140
[k]kxx=122
[k]ka=162
[k]kA=122
[k]ko=177
[k]koo=98
[k]ku=104
[k]kuu=122
[g]=89
```

```
[8]8=101
[8]8e=67
[8]8E=101
[8]8i=108
[8]8ii=101
[8]8w=101
[8]8ww=101
[8]8x=101
[8]8xx=101
[8]8a=118
[8]8A=110
[8]8o=143
[8]8oo=101
[8]8u=68
[8]8uu=92
[h]=97
[h]e=62
[h]E=97
[h]i=157
[h]ii=133
[h]w=59
[h]ww=91
[h]x=109
[h]xx=106
[h]a=86
[h]A=99
[h]o=102
[h]oo=86
[h]u=69
[h]uu=102
[h]h=97
[h]he=132
[h]hE=97
[h]hi=37
[h]hii=61
[h]hw=135
[h]hww=103
[h]hx=85
[h]hxx=88
[h]ha=108
[h]hA=95
[h]ho=92
[h]hoo=108
[h]hu=125
[h]huu=92
[j]=75
[j]e=60
[j]E=75
[j]i=75
[j]ii=75
[j]w=78
[j]ww=65
```

```
[t]6u=78
t[6]u=77
[t]6uu=87
t[6]uu=86
[t]t6=110
t[t]6=110
tt[6]=110
[t]t6e=110
t[t]6e=110
tt[6]e=110
[t]t6E=75
t[t]6E=75
tt[6]E=75
[t]t6i=95
t[t]6i=95
tt[6]i=95
[t]t6ii=110
t[t]6ii=110
tt[6]ii=110
[t]t6w=98
t[t]6w=98
tt[6]w=98
[t]t6x=98
t[t]6x=98
tt[6]x=98
[t]t6a=98
t[t]6a=98
tt[6]a=98
[t]t6o=98
t[t]6o=98
tt[6]o=98
[t]t6oo=86
t[t]6oo=86
tt[6]oo=86
[t]t6u=112
t[t]6u=112
tt[6]u=112
[t]t6uu=98
t[t]6uu=98
tt[6]uu=98
[d]z=113
d[z]=75
[d]zE=100
d[z]E=68
[d]zi=120
d[z]i=79
[d]zii=113
d[z]ii=75
[d]za=120
d[z]a=78
[d]zA=113
d[z]A=75
```

```
[w]wf=51
[x]f=81
[x]xf=60
[a]f=140
[A]f=250
[o]f=110
[o]of=64
[u]f=102
[u]uf=73
[e]v=98
[E]v=235
[i]v=110
[i]iv=71
[w]v=133
[w]wv=63
[x]v=83
[x]xv=57
[a]v=121
[A]v=243
[o]v=102
[o]ov=110
[u]v=90
[u]uv=153
[e]6=98
[E]6=153
[i]6=68
[i]i6=136
[w]6=125
[w]w6=150
[x]6=90
[x]x6=43
[a]6=143
[A]6=166
[o]6=110
[o]o6=31
[u]6=60
[u]u6=200
[e]z=121
[E]z=259
[i]z=157
[i]iz=32
[w]z=98
[w]wz=157
[x]z=184
[x]xz=91
[a]z=121
[A]z=174
[o]z=92
[o]oz=74
[u]z=188
[u]uz=94
[e]s=90
[a]tt6=117
[A]tt6=235
[o]tt6=149
[o]ott6=63
[u]tt6=141
[u]utt6=102
[e]ddz=109
[E]ddz=178
[i]ddz=106
[i]iddz=87
[w]ddz=111
[w]wddz=102
[x]ddz=105
[x]xddz=83
[a]ddz=110
[A]ddz=196
[o]ddz=107
[o]oddz=40
[u]ddz=109
[u]uddz=94
[e]tts=90
[E]tts=178
[i]tts=77
[i]itts=116
[w]tts=111
[w]wtts=124
[x]tts=105
[x]xtts=83
[a]tts=125
[A]tts=196
[o]tts=129
[o]otts=28
[u]tts=110
[u]utts=102
[e]dd8=109
[E]dd8=178
[i]dd8=106
[i]idd8=87
[w]dd8=111
[w]wdd8=102
[x]dd8=117
[x]xdd8=71
[a]dd8=143
[A]dd8=196
[o]dd8=121
[o]odd8=49
[u]dd8=109
[u]udd8=94
[e]7=82
[E]7=178
[i]7=62
[i]i7=131
[g]e=98
[g]E=87
[g]i=141
[g]ii=89
[g]w=78
[g]ww=75
[g]x=94
[g]xx=89
[g]a=64
[g]A=86
[g]o=75
[g]oo=90
[g]u=89
[g]uu=89
[g]g=103
[g]ge=106
[g]gE=140
[g]gi=71
[g]gii=103
[g]gw=110
[g]gww=117
[g]gx=98
[g]gxx=103
[g]ga=55
[g]gA=118
[g]go=117
[g]goo=102
[g]gu=103
[g]guu=103
[f]=129
[f]e=140
[f]E=129
[f]i=98
[f]ii=129
[f]w=121
[f]ww=128
[f]x=129
[f]xx=129
[f]a=121
[f]A=141
[f]o=153
[f]oo=129
[f]u=129
[f]uu=129
[f]f=119
[f]fe=103
[f]fE=119
[f]fi=74
[f]fii=119
[f]fw=169
[f]fww=120
[f]fx=119
[j]x=75
[j]xx=75
[j]a=68
[j]A=105
[j]o=75
[j]oo=94
[j]u=55
[j]uu=75
[j]j=75
[j]je=90
[j]jE=75
[j]ji=75
[j]jii=75
[j]jw=72
[j]jww=85
[j]jx=75
[j]jxx=75
[j]ja=82
[j]jA=45
[j]jo=75
[j]joo=56
[j]ju=95
[j]juu=75
[l]=57
[l]e=52
[l]E=64
[l]i=68
[l]ii=57
[l]w=43
[l]ww=52
[l]x=57
[l]xx=57
[l]a=37
[l]A=86
[l]o=57
[l]oo=62
[l]u=45
[l]uu=57
[l]l=80
[l]le=72
[l]lE=41
[l]li=69
[l]lii=80
[l]lw=94
[l]lww=88
[l]lx=80
[l]lxx=37
[l]la=91
[l]lA=86
[l]lo=80
[l]loo=134
[l]lu=92
[d]dz=100
d[d]z=65
dd[z]=110
[t]s=76
t[s]=75
[t]sE=64
t[s]E=64
[t]si=83
t[s]i=82
[t]sii=114
t[s]ii=113
[t]sw=61
t[s]w=60
[t]sww=76
t[s]ww=75
[t]sa=57
t[s]a=56
[t]sA=76
t[s]A=75
[t]so=55
t[s]o=55
[t]soo=76
t[s]oo=75
[t]suu=98
t[s]uu=98
[t]ts=91
t[t]s=90
tt[s]=67
[t]tse=90
t[t]se=90
tt[s]e=66
[t]tsE=91
t[t]sE=90
tt[s]E=67
[d]8=107
d[8]=107
[d]8e=69
d[8]e=68
[d]8E=153
d[8]E=153
[d]8A=99
d[8]A=99
[d]d8=75
d[d]8=75
dd[8]=100
[d]d8E=105
d[d]8E=105
dd[8]E=140
[7]=125
[7]e=137
[7]E=125
[7]i=125
```

```
[E]s=151
[i]s=105
[i]is=85
[w]s=98
[w]ws=90
[x]s=60
[x]xs=183
[a]s=98
[A]s=174
[o]s=98
[o]os=53
[u]s=105
[u]us=99
[e]8=71
[E]8=141
[i]8=74
[i]i8=98
[w]8=94
[w]w8=283
[x]8=84
[x]x8=126
[a]8=136
[A]8=214
[o]8=180
[o]o8=24
[u]8=115
[u]u8=93
[e]h=141
[E]h=158
[i]h=86
[i]ih=51
[w]h=141
[w]wh=24
[x]h=117
[x]xh=103
[a]h=90
[A]h=281
[o]h=90
[o]oh=84
[u]h=113
[u]uh=134
[e]j=149
[E]j=165
[i]j=105
[i]ij=49
[w]j=94
[w]wj=38
[x]j=86
[x]xj=64
[a]j=113
[A]j=158
[o]j=104
[w]7=113
[w]w7=91
[x]7=98
[x]x7=90
[a]7=115
[A]7=193
[o]7=113
[o]o7=87
[u]7=90
[u]u7=55
[e]4=78
[E]4=150
[i]4=150
[i]i4=117
[w]4=90
[w]w4=232
[x]4=90
[x]x4=93
[a]4=62
[A]4=193
[o]4=127
[o]o4=58
[u]4=133
[u]u4=134
[e]9=121
[E]9=251
[i]9=94
[i]i9=130
[w]9=98
[w]w9=76
[x]9=150
[x]x9=50
[a]9=75
[A]9=187
[o]9=99
[o]o9=26
[u]9=75
[u]u9=70
[p]=133
[p]e=104
[p]E=189
[p]i=113
[p]ii=133
[p]w=94
[p]ww=133
[p]x=133
[p]xx=133
[p]a=121
[p]A=165
[p]o=143
[p]oo=133
[p]u=133
[f]fa=127
[f]fA=145
[f]fo=95
[f]foo=119
[f]fu=119
[v]=78
[v]e=47
[v]E=58
[v]i=78
[v]w=78
[v]x=78
[v]a=70
[v]A=98
[v]o=78
[v]oo=110
[v]u=86
[v]uu=78
[v]v=105
[v]ve=145
[v]vE=125
[v]vi=105
[v]vii=98
[v]vw=105
[v]vww=105
[v]vx=105
[v]vxx=105
[v]va=110
[v]vA=86
[v]vo=105
[v]voo=73
[v]vu=97
[v]vuu=105
[6]=141
[6]e=105
[6]E=141
[6]i=200
[6]ii=141
[6]w=136
[6]ww=141
[6]x=141
[6]xx=141
[6]a=121
[6]A=157
[6]o=118
[6]oo=151
[6]u=141
[6]uu=141
[6]6=118
[6]6e=112
[6]6E=161
[6]6i=106
[6]6ii=118
[l]luu=80
[r]=38
[r]e=35
[r]E=26
[r]i=58
[r]ii=47
[r]w=37
[r]ww=35
[r]x=30
[r]xx=52
[r]a=30
[r]A=30
[r]o=31
[r]oo=38
[r]u=47
[r]uu=38
[r]r=88
[r]re=91
[r]rE=91
[r]ri=68
[r]rii=102
[r]rw=89
[r]rww=91
[r]rx=96
[r]rxx=74
[r]ra=69
[r]rA=80
[r]ro=126
[r]roo=88
[r]ru=79
[r]ruu=88
[n]=63
[n]e=52
[n]E=86
[n]i=68
[n]ii=63
[n]w=62
[n]ww=75
[n]x=75
[n]xx=63
[n]a=52
[n]A=41
[n]o=55
[n]oo=62
[n]u=63
[n]uu=63
[n]n=127
[n]ne=120
[n]nE=118
[n]ni=152
[n]nii=127
[n]nw=118
[7]ii=125
[7]w=110
[7]ww=125
[7]x=125
[7]xx=142
[7]a=111
[7]A=125
[7]o=125
[7]oo=125
[7]u=125
[7]uu=125
[7]7=129
[7]7e=117
[7]7E=150
[7]7i=129
[7]7ii=129
[7]7w=144
[7]7ww=129
[7]7x=129
[7]7xx=112
[7]7a=122
[7]7A=102
[7]7o=129
[7]7oo=129
[7]7u=118
[7]7uu=165
[4]=89
[4]e=56
[4]E=89
[4]i=89
[4]ii=89
[4]w=83
[4]ww=89
[4]x=98
[4]xx=113
[4]a=90
[4]A=103
[4]o=89
[4]oo=78
[4]u=89
[4]uu=89
[4]4=175
[4]4e=223
[4]4E=175
[4]4i=175
[4]4ii=175
[4]4w=181
[4]4ww=175
[4]4x=166
[4]4xx=151
[4]4a=174
[4]4A=124
```

```
[o]oj=98
[u]j=150
[u]uj=46
[e]l=142
[E]l=200
[i]l=102
[i]il=87
[w]l=70
[w]wl=63
[x]l=157
[x]xl=118
[a]l=102
[A]l=151
[o]l=136
[o]ol=29
[u]l=90
[u]ul=145
[e]r=90
[E]r=166
[i]r=145
[i]ir=74
[w]r=125
[w]wr=165
[x]r=117
[x]xr=73
[a]r=117
[A]r=205
[o]r=93
[o]or=96
[u]r=95
[u]ur=115
[e]m=98
[E]m=174
[i]m=133
[i]im=87
[p]uu=133
[p]p=125
[p]pe=202
[p]pE=101
[p]pi=138
[p]pii=125
[p]pw=181
[p]pww=125
[p]px=125
[p]pxx=125
[p]pa=60
[p]pA=78
[p]po=116
[p]poo=125
[p]pu=125
[p]puu=125
[b]=85
[b]e=75
[b]E=94
[b]i=83
[b]ii=85
[b]w=60
[b]ww=85
[b]x=68
[b]xx=85
[b]a=113
[b]A=85
[b]o=77
[b]oo=110
[b]u=85
[b]uu=85
[b]b=129
[b]be=149
[b]bE=126
[b]bi=160
[6]6w=123
[6]6ww=118
[6]6x=118
[6]6xx=185
[6]6a=138
[6]6A=67
[6]6o=93
[6]6oo=108
[6]6u=118
[6]6uu=86
[z]=89
[z]e=64
[z]E=89
[z]i=68
[z]ii=125
[z]w=89
[z]ww=89
[z]x=83
[z]xx=89
[z]a=86
[z]A=89
[z]o=71
[z]oo=90
[z]u=133
[z]uu=82
[z]z=112
[z]ze=107
[z]zE=138
[z]zi=108
[z]zii=102
[z]zw=112
[z]zww=112
[z]zx=118
[z]zxx=112
[z]za=134
[n]nww=204
[n]nx=115
[n]nxx=67
[n]na=84
[n]nA=102
[n]no=135
[n]noo=162
[n]nu=145
[n]nuu=127
[m]=80
[m]e=52
[m]E=94
[m]i=80
[m]ii=94
[m]w=73
[m]ww=80
[m]x=80
[m]xx=80
[m]a=71
[m]A=80
[m]o=68
[m]oo=110
[m]u=77
[m]uu=80
[m]m=102
[m]me=113
[m]mE=88
[m]mi=81
[m]mii=98
[m]mw=123
[m]mww=102
[m]mx=102
[m]mxx=102
[m]ma=113
[m]mA=104
[4]4o=175
[4]4oo=189
[4]4u=175
[4]4uu=193
[9]=63
[9]e=52
[9]E=86
[9]i=68
[9]ii=63
[9]w=62
[9]ww=75
[9]x=75
[9]xx=63
[9]a=52
[9]A=41
[9]o=55
[9]oo=62
[9]u=63
[9]uu=63
[9]9=127
[9]9e=120
[9]9E=118
[9]9i=152
[9]9ii=127
[9]9w=118
[9]9ww=204
[9]9x=115
[9]9xx=67
[9]9a=84
[9]9A=102
[9]9o=135
[9]9oo=162
[9]9u=145
[9]9uu=127
```

Bibliography

Abramson, N.: Information theory and coding. New York 1963.

Attneave, F.: Some informational aspects of visual perception. Psychological Review 61-3, 1954, p.183-193.

Baddeley, A.D., Thomson, N. and Buchanan, M.: Word length and the structure of short-term memory. Journal of Verbal Learning and Verbal Behaviour, 14, 1975, p.575-589.

Baddeley, A.D.: Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. Quarterly Journal of Experimental Psychology, 18, 1966(b), p.362-365.

Baddeley, A.D.: The influence of acoustic and semantic similarity on long-term memory for word sequences. Quarterly Journal of Experimental Psychology, 18, 1966(a), p.302-309.

Baddeley, A.D.: Working memory. Oxford 1986.

Ben-Zeev, S.: The effect of Spanish-English in children from less priviledged neighborhoods on cognitive development and cognitive strategy. Working papers on bilingualism 14, 1977(b), p.83-122.

Ben-Zeev, S.: The influence of biligualism on cognitive development and cognitive strategy. Child Development 48, 1977(a), 1009-1018.

Benveniste, E.: Nature du signe linguistique. Acta Linguistica I, 1939.

Bolla, K.: Magyar hangalbum. In: Bolla, K.(ed): Fejezetek a magyar leíró hangtanból. Budapest 1982.

Brown, J.A.: Some tests on the decay theory of immediate memory. Quarterly Journal of Experimental Psychology, 10, 1958, p.12-21.

Bybee, J.: Morphology. A study of the relation between meaning and form. Amsterdam/Philadelphia 1985.

Chomsky, N.: Syntactic structures. The Hague 1957.

Chomsky, N.: Three models for the description of language. I.R.E. transactions on information theory, vol.IT-2, Proceedings of the symposium on information theory, 1956.

Coombs, C.H., Dawes, R.N. and Tversky, A.: Mathematical psychology. Englewood Cliffs, New Jersey 1970.

Cummins, J. and Swain, M.: Bilingualism in education. London 1986.

Cummins, J.: The cross-lingual dimensions of language proficiency: implications for bilingual education and the optimal age issue. TESOL Quarterly 14, 1980, p. 175-187.

Ellis, H.C. and Hunt, R.R.: Fundamentals of human memory and cognition. Dubuque, Iowa 1989.

Estes, W.: Towards a statistical theory of learning. Psychological Review, 57, 1950, p.94-107.

Eysenck, M.W. and Keane, M.T.: Cognitive psychology. London 1990.

Eysenck, M.W.: Depth, elaboration and distinctiveness. In: L.S.Cermak and F.I.M.Craik (eds.): Levels of processing in human memory. Hillsdale, N.J. 1979.

Füredi M. and Kelemen, J.: A mai magyar nyelv szépprózai gyakorisági szótára. Budapest 1989.

Garner, W.R.: The processing of information and structure. New York 1974.

Garner, W.R.: Uncertainty and structure as psychological concepts. New York 1962.

Garside,R., Leech,G. and Sampson,G.: The computational analysis of English. London 1987.

Gósy, M.: Beszédészlelés. Budapest 1989.

Harris, Z.: Distributional structure. Word 10, 1954.

Haslerud,G.M. and Clark, R.E.: On the reintegrative perception of words. American Journal of Psychology, 70, 1957, p.97-101.

Ianco-Worall, A.: Biligualism and cognitive development. Child Development 43, 1972, p.1390-1400.

Kassai, I.: A magyar beszédhangok idôtartamviszonyai. In: Bolla, K.(ed): Fejezetek a magyar leíró hangtanból. Budapest 1982.

Kiss, K.É.: Configurationality in Hungarian. Budapest 1987.

Krippendorff, K.: Information theory. London 1986.

Lakoff, G.: Women, fire and dangerous things. Chicago 1987.

Lavotha, Ö. and Lavotha, Cs.: Ungersk grammatik. Stockholm 1973.

Lyons, J.: Semantics. Cambridge 1977.

Magdics, K.: Studies in the acoustic characteristics of Hungarian speech sounds. The Hague 1969.

Mandelbrot, B.: Structure formelle des textes et communication. Word 10, 1954.

McShane, J.: Cognitive development. Oxford 1991.

Márai, S.: Egy polgár vallomásai. Budapest 1990.

Peterson, L.R. and Peterson, M.J.: Short-term retention of individual verbal terms. Journal of Experimental Psychcology, 58, 1959, p.193-198.

Rivera, C.: Language proficiency and academic achievement. Multilingual Matters 10, 1984.

Rumelhart, D.E. and McClelland, J.R.: An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. Psychological Review 89-1, 1982, p.60-94.

Sampson, G.: Probabilistic parsing. Reprint paper for the Nobel Symposium on Corpus Linguistics. Stockholm 1991.

Saussure, F.: Cours de linguistique générale. Édition critique préparée par Tullio de Mauro. Paris 1972.

Shannon, C.A. and Weaver, W.: The mathematical theory of communication. Urbana 1948.

Taylor, J.: Linguistic categorization. Oxford 1989.

Tompa, J.: Kleine Ungarische Grammatik. Budapest 1972.

Tompa, J.: Ungarische Grammatik. Budapest 1968.

Treisman, A.M.: Verbal cues, language and meaning in selective attention. American Journal of Psychology, 77, 1964, p.206-219.

Tulving, E.: Relations between encoding specificity and levels of processing. In: L.S. Cermak and F.I.M.Craik (Eds.): Levels of processing in human memory. Hillsdale, N.J. 1979.

Wickelgren, W.A.: Phonetic coding and serial order. In: Carterette, E.C. and Friedman, M.P.: Handbook of perception. Vol.VII: Language and speech. New York 1976.

Wiseman, S. and Tulving, E.: Encoding specificity. Relation between recall superiority and recognition failure. Journal of experimental psychology: Human Learning and Meaning, 2, 1976, p.349-361.