This is a translation into English of an article originally published in Dutch, as "Taaltheorie en taaltechnologie; competence en performance", in: R. de Kort and G.L.J. Leerdam (eds.): *Computertoepassingen in de Neerlandistiek*. Almere: LVVN, 1990, pp. 7-22.

Remko Scha
Institute for Logic, Language and Computation
University of Amsterdam

Language theory and language technology; competence and performance.

Summary

The current generation of language processing systems is based on linguistically motivated competence models of natural languages. The problems encountered with these systems suggest the need for performance-models of language processing, which take into account the statistical properties of actual language use. This article describes the overall set-up of such a model. The system I propose employs an annotated corpus; in analysing new input it tries to find the most probable way to reconstruct this input from fragments that are already contained in the corpus. This perspective on language processing also has interesting consequences for linguistic theory; some of these are briefly discussed.

1. Introduction.

The starting point for this article was the question: what significance can language technology have for language theory? The usual answer to this question is, that the application of the methods and insights of theoretical linguistics in working computer programs is a good way to test and refine these theoretical ideas. I agree with this answer, and I will emphatically reiterate it here. But most of this article is devoted to a somewhat more speculative train of thought which shows that language-technological considerations can have important theoretical implications.

My considerations focus on a fundamental problem which is faced by current language-processing systems: the problem of ambiguity. To solve the ambiguity problem it is necessary to put linguistic insights about the structure and meaning of language utterances under a common denominator with statistical data about actual language use. I will sketch a technique which might be able to do this: data-oriented parsing, by means of pattern-matching with an annotated corpus. This parsing technique may be of more than technological interest: it suggests a new and attractive perspective on language and the language faculty.

First a warning. The following discussion concentrates almost exclusively on the problem of syntactic analysis. Of course this is only a sub-problem -- both in language theory and in language technology. But this problem turns out to yield so much food for thought already, that it does not seem useful to complicate the discussion by addressing the integration with phonetics, phonology, morphology, semantics, pragmatics and discourse-processing. How the different kinds of linguistic knowledge in a language-processing system ought to be distributed over the modules of the algorithm, is a question which will be left out of consideration completely.

2. Linguistics and language technology.

To be able to turn linguistics into a hard science, Chomsky [1957] assigned a mathematical correlate to the intuitive idea of a "language". He proposed to identify a language with a set of sentences: with the set of grammatically correct utterance forms that are possible in the language. The goal of descriptive linguistics is then to characterise, for individual languages, the set of grammatical sentences explicitly, by means of a formal grammar. And the goal of explanatory linguistic theories should then be, to determine the universal properties which the grammars of all languages share, and to give a psychological account of these universals.

In this view, linguistic theory is not immediately concerned with describing the actual language use in a language community. Although we may assume that there is a relation between the language users' grammaticality intuitions and their actual language behaviour, we must make a sharp distinction between these; on the one hand the language system may offer possibilities which are rarely or never used; on the other hand the actual language use involves mistakes and sloppinesses which a linguistic theory should not necessarily account for. In Chomsky's terminology: linguistics is concerned with the linguistic *competence* rather than the actual *performance* of the language user. Or, in the words of Saussure, who had emphasized this distinction before: with *langue* rather than *parole*.

Chomsky's work has constituted the methodological paradigm for almost all linguistic theory of the last few decades. This comprises not only the research tradition that is explicitly aiming at working out Chomsky's syntactic insights. The perspective summarized above has also determined the goals and methods of the most important alternative approaches to syntax, and of the semantic research traditions which have grown out of Richard Montague's work. Now we may ask: how does language technology relate to this language-theoretical paradigm?

Relatively few language technologists invoke Chomsky's ideas explicitly; but their methodological aassumptions tend to be implicitly based on his paradigm. Of course there are also important differences between the theoretically oriented and the technologically oriented language research. Compared to theoretical linguistics, language-technological research has usually been more descriptive, and less concerned with the universal validity and the explanatory power of the theory. In developing a translation system or a natural-language database-interface, the descriptive adequacy of the grammar of the input language has obviously a higher priority than gaining insights about syntactic universals. Equally evident is the observation that the syntactic and semantic rules developed for a language-technological application must be articulated in a strictly formal way, whereas the results of theoretical research may often take the form of essayistic reflections on different variations of an informally presented idea.

We thus see a complementary relation between theoretical linguistics and language technology; the theory is concerned, often in an informal way, with the general structure of linguistic competence and Universal Grammar; in language technology one tries to specify, in complete formal detail, descriptively adequate grammars of individual languages. Therefore, language-technological work will eventually be of considerable theoretical importance: the theoretical speculations about the structure of linguistic competence can only be validated if they give rise to a formal framework which allows for the specification of descriptively adequate grammars. Because theoretical linguists do not seem particularly interested in this boundary condition of their work, the application-oriented grammar-development activities constitute a useful and necessary complement to theoretical linguistic research.

Language-technological work has shown in the meantime that for the development of theoretically interesting grammars, computational support is indispensible. Formal grammars which describe some non-trivial phenomena in a partially correct way tend to get extremely complex -- so complex, that it is difficult to imagine how they could be tested, maintained and extended without computational tools.

There is another reason why language technology is interesting for linguistic theory: language-technological applications involve systems which are intended to work with some form of "real language" as input. Implementing a competence grammar will therefore be not enough in the end: one also needs software which deals with possibly relevant performance phenomena, and this software must interface in an adequate way with the competence grammar. The possibility of complementing a competence-grammar with an account of performance phenomena is another boundary condition of current linguistic theory which does not receive a lot of attention in theoretical research. Language-technological research may also be of theoretical importance here.

There are thus many opportunities for interesting interactions between language theory and language technology; but until recently such interactions did not often occur. For a long time, language technology has developed in relative isolation from theoretical linguistics. This isolation came about because Chomsky's formulation of his syntactic insights crucially used the notion of a "transformation" -- and many found this a computationally unattractive notion, especially for analysis-algorithms. Computational linguists felt they needed to develop alternative methods for language description which were more directly coupled to locally observable properties of *surface structure*, and therefore more easy to implement; this gave rise to Augmented Transition Networks and enriched contextfree grammars. After the heydays of Transformational Grammar were over, there has been a remarkable *rapprochement* between language theory and language technology, because enriched contextfree grammars, which are considered computationally attractive, acquired theoretical respectability. Gazdar, Pullum and Sag created a breakthrough in this area with their Generalized Phrase Structure Grammar.

For enriched contextfree grammars, effective parsing algorithms have been developed. There are procedures which establish the grammaticality of an arbitrary input-sentence in a reasonably efficient way, and determine its structural description(s) as defined by the grammar. This has made it possible to implement interesting prototype-systems which analyse their input in accordance with such a grammar. The results of this approach have been encouraging. They were certainly better than the results of competing approaches from Artificial Intelligence which worked without formal syntax (such as the prototypical versions of "frame-based parsing" and "neural networks"). Nevertheless the practical application of linguistic grammars in language processing systems is not without problems. These we consider in the next section.

3. Limitations of current language processing systems.

The applicability of currently existing linguistic technology depends of course on the availability of descriptively adequate grammars for substantial fragments of natural languages. But writing a system of rules which provides a good characterization of the grammatical structures of a natural language turns out to be surprisingly difficult. There is no formal grammar yet which correctly describes the richness of a natural language -- not even a formal grammar which adequately covers a non-trivial corpus of a substantial size. The problem is not only that the grammar of natural language is large and complex, and that we therefore still need hard work and deep thought to describe it. The process of developing a formal grammar for a particular natural language is especially disappointing because it becomes increasingly difficult and laborious as the grammar gets larger. The larger the number of phenomena that are already partially accounted for, the larger the number of interactions that must be inspected when one tries to introduce an account of new phenomena.

A second problem with the current syntax/parsing-paradigm is even easier to notice: the problem of ambiguity. It turns out that as soon as a grammar characterises a non-trivial part of natural language, almost every input-sentence with a certain length has many (often very many) different structural analyses (and corresponding semantic interpretations). This is problematic because usually most of these interpretations are not perceived as possible by a human language user, although there is no reason to exclude them on formal syntactic or semantic grounds. Often it is only a matter of *relative*

implausibility: the only reason why the language user does not become aware of a particular interpretation of a sentence, is that another interpretation is *more* plausible.

The two problems I mentioned are not independent of each other. Because of the first problem (the disquieting combinatorics of interacting syntactic phenomena), we might be inclined to stop refining the syntactic subcategories at a certain point, thus ending up with a more "tolerant" grammar which accepts various less happy constructions as nevertheless grammatical. This is a possible strategy, because the Chomskyan paradigm does not clearly fix how language competence is to be delimited with respect to language performance. Not all judgments of sentences as "strange", "unusual", "infelicitous", "incorrect", or "uninterpretable" need to be viewed as negative grammaticality-judgments; ultimately the elegance of the resulting theory determines whether certain unwellformedness-judgments are to be explained by the competence-grammar or by the performance-module. But the designer of a language-processing system who relaxes the system's grammar is not finished by doing that: he is confronted with an increased ambiguity in the grammatical analysis process, and must design a performance-module which can make a sensible selection from the set of alternative analyses.

4. Competence and Performance.

The limitations of current language processing systems are not surprising: they follow immediately from the fact that these systems are built on a *competence-grammar* in the Chomskyan sense. As mentioned above, Chomsky made an emphatic distinction between the "competence" of a language user and the "performance" of this language user. The competence consists in the knowledge of language which the language user *in principle* has; the performance is the result of the psychological process that employs this knowledge (in producing or in interpreting language utterances).

The formal grammars that theoretical linguistics is concerned with, aim at characterising the competence of the language user. But the preferences that language users display in dealing with syntactically ambiguous sentences constitute a prototypical example of a phenomenon that in the Chomskyan view belongs to the realm of performance.

The ambiguity-problem discussed above follows from an intrinsic limitation of linguistic competence-grammars: such grammars define the sentences of a language and the corresponding structural analyses, but they do not specify a probability ordering or any other ranking between the different sentences or between the different analyses of one sentence. This limitation is even more serious when a grammar is used for processing input which frequently contains mistakes. Such a situation occurs in processing spoken language. The output of a speech recognition system is always very imperfect, because such a system often only makes guesses about the identity of its input-words. In this situation the parsing mechanism has an additional task, which it doesn't have in dealing with correctly typed alpha-numeric input. The speech recognition module may discern several alternative word sequences in the input signal; only one of these is correct, and the parsing-module must employ its syntactic information to arrive at an optimal decision about the nature of the input. A simple yes/no judgment about the grammaticality of a word sequence is insufficient for this purpose: many word sequences are strictly speaking grammatical but very implausible; and the number of word sequences of this kind gets larger when a grammar accounts for a larger number of phenomena.

To construct effective language processing systems, we must therefore implement performance-grammars rather than competence-grammars. These performance-grammars must not only contain information about the structural possibilities of the general language system, but also about "accidental" details of the actual language use in a language community, which determine the language experiences of an individual, and thereby influence what kind of utterances this individual expects to encounter, and what structures and meanings these utterances are expected to have.

The linguistic perspective on performance involves the implicit assumption that language behaviour can be accounted for by a system that comprises a competence-grammar as an identifiable sub-component. But because of the ambiguity problem this assumption is computationally unattractive: if we would find criteria to prefer certain syntactic analyses above others, the efficiency of the whole process might benefit if these criteria were applied in an early stage, integrated with the strictly syntactic rules. This would amount to an integrated implementation of competence- and performance-notions.

But we can also go one step further, and fundamentally question the customary concept of a competence-grammar. We can try to account for language-performance without invoking an explicit competence-grammar. (This would mean that grammaticality-judgments are to be accounted for as performance phenomena which do not have a different cognitive status than other performance phenomena.) This is the idea that I want to work out somewhat more concretely now. Later (in section 7) I will return to the possible theoretical merits of this point of view.

5. Statistics.

There is an alternative language description tradition which has always focussed on the concrete details of actual language use, often without paying much attention to the abstract language system: the *statistical* tradition. In this approach the characterisation of syntactic structures is often completely ignored; one only describes "superficial" statistical properties of representative language corpus that is as large as possible. Usually one simply indicates the occurrence frequencies of different words, the probability that a specific word is followed by another specific word, the probability that a specific sequence of 2 words is followed by a specific word, etc. (nth order Markov chains). See, for instance, Bahl et al. (1983), Jelinek (1986).

The Markov-approach has been very succesful for the purpose of selecting the most probable sentence from the set of possible outputs generated by a speech recognition component. It is clear, however, that

for various other purposes this approach is completely insufficient, beacuse it does not employ a notion of syntactic *structure*. For a natural-language database-interface, for instance, semantic interpretation rules must be applied, on the basis of a structural analysis of the input. And there are also statistically significant regularities in corpus-sentences which encompass long word sequences through syntactic structures; in the Markov approach these are ignored. The challenge is now, to develop a method for language description and parsing which does justice to the statistical as well as the structural aspects of language.

The idea that a synthesis between the syntactical and the statistical approaches would be useful and interesting has been broached incidentally before, but so far it has not been thought through very well. The only existing technical instantiation, the concept of a *stochastic grammar*, is rather simplistic. Such a grammar just juxtaposes the most basic syntactic notion with the most basic probabilistic notion: an "old-fashioned" contextfree grammar describes syntactic structures by means of a system of rewrite rules -- but the rewrite rules are provided with probabilities or with rankings that correlate with the application probabilities of the rules. (Derouault en Merialdo, 1986; Fusijaki, 1984; Fusijaki et al., 1989.)

A stochastic grammar which only assigns probabilities to individual syntactically motivated rewrite rules, cannot capture all relevant statistical properties of a language corpus. For instance, it cannot indicate how the occurrence probabilities of syntactic structures or lexical items depend on their syntactic/lexical context. As a result, it cannot even identify frequently occurring phrases and figures of speech -- a disappointing property, because one would like to see that the analyses constructed in terms of such phrases and figures of speech would automatically get a high priority in the ranking of the different possible syntactic analyses of a sentence.

6. A new approach: data-oriented parsing.

Current stochastic grammars operate with units that are too small: rewrite rules which describe exactly one level of the constituent structure of a sentence, and whose application probabilities are supposed to be context-independent. Instead, we would like to use the statistical approach while working with larger units.

There is in fact a linguistic tradition which has been thinking in this direction. Bolinger (1961, 1976), Becker (1984a, 1984b), and Hopper (1987) have distanced themselves emphatically from the usual formal grammars. They assign a central role to the concrete language data; they view new utterances as built up out of fragments culled from previously processed texts; idiomaticity is the rule rather than the exception.

These researchers have not put a big emphasis on formalising their ideas. Some of them even seem to consider their perspective on language intrinsically incompatible with formalisation; they concentrate completely on informal, anecdotal descriptions of very specific language phenomena, such as semi-idiomatic expressions and conventional turns of phrase. If we nevertheless want to try to work out these kinds of ideas in a formal way, we find the best point of departure in the work of Fillmore et al. (1988), who suggest to describe a language not by means of a set of rewrite rules, but by means of a set of "constructions". A construction is a tree-structure: a fragment of a constituent structure that can encompass more than one level. This tree is labelled with syntactic, semantic and pragmatic categories and feature-values. Lexical items can be specified as part of a construction. Constructions can be idiomatic: the meaning of a larger constituent can be specified without being built up out of the meanings of sub-constituents.

Fillmore's ideas betray the influence of the formal grammar tradition: the combinatorics of fitting constructions together, defines a class of sentences in a manner which strongly resembles a contextfree grammar. The way in which Fillmore generalizes the grammar notion, however, solves exactly the problem that we encountered in current stochastic grammars: if a "construction grammar" is combined with statistical notions, it may be possible to represent all relevant statistical information. We will work out this idea somewhat further now.

The human language interpretation process has a strong preference for recognizing sentences, phrases and patterns that have occurred before. Structures and interpretations which have occurred frequently are preferred above alternatives which have not or rarely been experienced before. All lexical elements, syntactic structures and "constructions" which the language user has ever encountered, and their frequency of occurrence, can have an influence on the processing of new input. The amount of information that is necessary for a realistic performance-model is therefore much larger than the grammars that we are used to. The language-experience of an adult language user consists of a large number of utterances. And every utterance contains a multitude of constructions: not only the whole sentence, and all its constituents, but also all patterns that we can abstract from these by substituting "free variables" for lexical elements or complex constituents.

How can we represent all this information, annotated with occurrence frequencies, for a large corpus? Unlike the human brain, currently available digital storage-media are not built for immediate and flexible associative access to gigantic databases. Before we can use this approach to construct practically useful systems, considerable implementation problems must therefore be solved. But if we leave the implementation problems aside, and focus on *what* is to be implemented, the situation is quite simple. The information that is needed, is a maximally adequate representation of the concrete past language experience of the language user. That is: a corpus that is as large as possible, containing utterances with their syntactic analyses and semantic interpretations. Because *all* occurring patterns, and their frequencies, can have an influence on the processing of new input, there is hardly any information in the corpus that can be ignored. Conceptually we might as well suppose that the language processing system has access to the whole corpus.

If we indeed assume that a person's language processing system has access to this person's complete

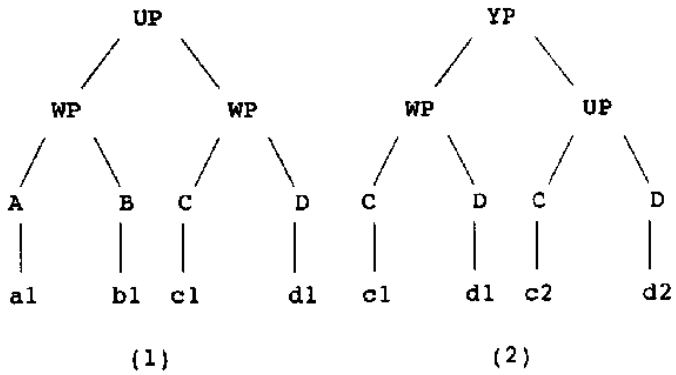
past language experience (or a sufficiently large sample of it), as represented by a corpus of sentences with syntactic analysis trees and semantic interpretations, a new perspective on parsing becomes possible. Parsing does not have to happen by applying grammatical rules to the input sentence now; it may be a matching process that attempts to construct an optimal analogy between the input-sentence and as many corpus-sentences as possible. The nature of this process can best be illustrated by looking at two extreme cases. On the one hand a sentence may be recognized because it occurs literally in the corpus; preferably, this sentence then gets the same analysis as the corresponding corpus-sentence. On the other hand, the matching process may have to take many different corpus-sentences into consideration, and may have to abstract away from most of the properties of each of them; in such a case, the system is in fact extracting the relevant grammar rules on the fly from the corpus. Most input-sentences will lie somehere between these two extremes: certain abstractions are necessary to make a succesful match, but some combinations of the properties of the input sentence occurred already in the corpus and can be treated as one unit (a "construction" à la Fillmore) by the matching-process.

In developing the further details of matching-process, two boundary-conditions must be kept in mind:

(1) The analogy between the input and the corpus must preferably be constructed in the simplest possible way. The fragments and patterns from the input-sentence which are found in the corpus, should preferably cover a maximally large part of the constituent-structure. In other words: the number of constructions that is used to re-construct the sentence in order to recognize it must be as *small* as possible.

(2) The surprising results that have been achieved with simple statistical models indicate that the frequency of occurrence of constructions in the corpus must play a role in the analysis process. More frequent constructions are to be preferred above less frequent ones. (If we look at the process from a statistical point of view, we see this effect can be attained in an elegant, implicit way by searching the corpus "at random" for matching constructions!)

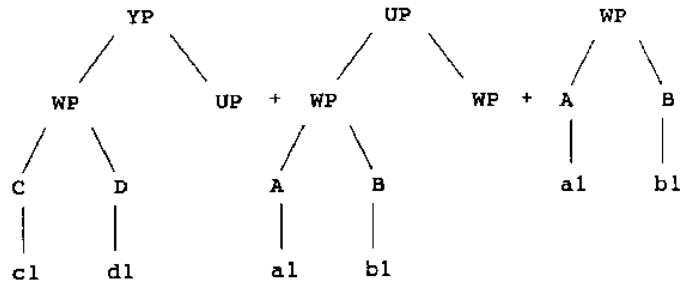
Finally an extremely simple formal example without any linguistic content, to demonstrate the basic idea. Assume the whole corpus consists of only these two trees:



The input string "c2 d2" can now be recognized as a UP by an exact match with the UP constituent in tree (2). It can also be recognized as a WP by combining the rule WP => C + D which can be recognized in (1) as well as in (2), with the rules C => c2 and D => d2 which can be found in (2). We deem the latter analysis less plausible, because its construction takes three steps.

In the same way there are two possibilities for the analysis of the string "c1 d1". It can be viewed as a WP because of exact recognition in (1) and (2), and as a UP by combining rewrite rules for UP, C and D. Because the exact recognition of "c1 d1" as a WP occurs twice in this corpus, and the rewrite rule UP => C + D only once, the preference for the analysis that takes place in one single recognition step is even more clear in this case.

A somewhat more complex analysis is illustrated by the string "c1 d1 a1 b1 a1 b1", which is parsed as a YP by combining the following constructions from the corpus:



which occur respectively in tree (2), tree (1) and tree (1). This analysis shows that constructions à la Fillmore can be applied recursively in the matching process, just like contextfree rules.

7. Theoretical consequences of this approach.

The parsing technology proposed here implies a perspective on language description that differs essentially from the usual one. I shall now discuss some interesting theoretical consequences of this new perspective.

1.

The distinction between grammatical and ungrammatical sentences looses its absolute character. At first sight this may seem unattractive, in a linguistic tradition that is founded on Chomsky's idea of a formal grammar as a recursive characterisation of exactly all grammatical sentences of a language. But in fact it ties in with a development that has been going on for some time in theoretical linguistic discussions. These discussions often concern interactions between complex phenomena, which are illustrated by sentences which are so complicated that even the professional linguist finds it difficult to pass absolute grammaticality judgments on them. Instead, one therefore assigns *relative* grammaticality judgments to such sentences: they are not categorised as grammatical or ungrammatical, but they are ordered relative to each other as being *more* or *less* grammatical. One thus assumes that a language does not have one grammar (in the Chomskyan sense), but a partially ordered multiplicity of grammars. This assumption tends to remain implicit; and most implemented systems still employ just one grammar.

I expect that in the process model proposed here, the most plausible sentences can be analysed with little effort, and that the analysis of more unusual and less grammatical sentences will take much more processing time. I hope that the processing model can be worked out in such a way that this "syntactic difficulty" ordering can be made to tally with linguistic judgments concerning relative grammaticality.

2.

A related phenomenon is the instability of grammaticality judgments. Whether a sentence is considered grammatical, often depends on the context in which it is presented -- on its semantic/pragmatic plausibility, but also on the structure of sentences that came before. Matthews (1979) argues that grammaticality may in fact not be a decidable property of sentences at all. He gives concrete examples of the instability of grammaticality judgments, showing how a sentence with a particular structure may enable a language user to perceive an analogous structure in a subsequent sentence, that would have been impossible to perceive without this priming. This phenomenon is absolutely incompatible with the standard parsing methods, that work exclusively with grammatical rules, but it fits very well into the data-oriented approach proposed here: if recent utterances are weighted more heavily in the matching process, we get precisely this kind of bias.

Our account of grammaticality judgments thus ties in with what Stich (1971) has suggested. Grammaticality judgments are not created by applying a precompiled set of grammar rules; rather, they constitute a perceptual judgment about the degree of similarity between the sentence under consideration and the sentences which the language user has stored as examples of grammaticality. A language user's concrete past language experiences determine how a new utterance is processed. There is no evidence for the assumption that these past language experiences are generalized into a consistent theory which defines the grammaticality and the structure of new utterances in an unequivocal fashion.

3.

Saussure's paradigm, which has completely dominated linguistic research in the last few decades, is called into question by the research proposed here. Jameson (1972) observed that "Saussure's originality was to have insisted on the fact that language as a total system is complete at every moment, no matter what happens to have been altered in it a moment before". The assumption of a consistent and complete system has been one of the pillars of Chomsky's syntactic tradition as well as Montague's semantic tradition. I would like to cast some doubt on this assumption. When we assign a central role to concrete language data, the consistent algebraic language system becomes an epiphenomenon which may turn out to be largely illusory. It is very well possible that the language system is a non-deterministic conglomerate of incompatible but overlapping "subsystems".

4.

Giving up Saussure's paradigm would create new possibilities for developing a formal account of the process of language change -- something which is hard to do when language is viewed as a mathematical system which at any moment is consistent and complete. "L'opposition synchronie/diachronie se situant à l'interieur de la langue, le changement est, chez F. de Saussure, le lieu d'un paradoxe: c'est par des actes de parole que la langue change". (Robert, 1977)

5.

Fodor (1975) unearthed a similar paradox concerning the language acquisition of the individual language user. If human linguistic cognition can at any moment be described as a consistent system that performs computations on mathematically well-defined "representations", it becomes quite mysterious how someone may ever learn a "really new" concept: all concepts that a person's thinking may ever employ, must already be definable in terms of the algebra of elementary concepts and operations of a person's "language of thought". On these assumptions, a person's conceptual repertoire must be completely "innate". Surprisingly, Fodor accepts this absurd conclusion.

The absurdity of Fodor's point of view has been widely recognized, and there even is a growing unanimity concerning the answer to the "Fodor-paradox": we must look at "subsymbolic processes". The underlying idea is, that the cognitive system does not "really" work with symbolic representations. The symbols are only "emergent phenomena".

So far, the idea of a subsymbolic account of language processing has been mostly worked out in terms of "connectionist" (neurologically inspired) models (cf. for instance Rumelhart et al. (1986), McClelland et al. (1986)). That does not do justice to this idea. The capabilities of connectionist

networks are limited; it is not clear yet to what extent they can carry out other tasks than simple classifications in n-dimensional space. The results are modest therefore, and are easy to criticize. (Cf. for instance Pinker & Mehler (1988), Levelt (1989).)

The connectionist research program conflates two distinct research goals. First of all one tries to work out some quite plausible ideas about the statistical, data-oriented character of language processing and other cognitive activities. Secondly, one is committed to implementing these statistical processeson a very specific kind of distributed hardware-architecture (connectionist networks). The underlying presumption is, that the operation of connectionist networks is a meaningful idealisation of the elementary processes which implement human cognition in the human brain; this presumption is highly questionable, however. Linguistic research which is committed to this interesting but very difficult implementation environment does not do justice, therefore, to the potential of the statistical perspective on cognition, which emphasizes data and perception rather than rules and reasoning.

It may be one of the most important merits of the approach sketched above, that it shares many properties with the connectionist approach, while assuming a more powerful and flexible computational framework. Like the proponents of neural networks, we try to explain the processing of a new utterance in terms of the integration of this utterance with the sum of all previously stored utterances; we avoid invoking explicitly stored abstract rules. But the matching process we envisage may access the linguistic structure of the stored utterances; and we do not impose any constraints on the nature and the complexity of this process.

8. Conclusions and further research.

Above we have shown that ideas which originate in the problems of language technology may be interesting for linguistic theory. We have not yet shown, however, that these ideas are indeed correct. A lot of research is still needed to work them out and validate them. I now briefly discuss some research questions which are raised here.

1.

First of all the matching algorithm that was impressionistically sketched above must be specified in detail, and its properties must be tested in practice. It will be especially interesting to see how such an algorithm may deal with complex syntactic phenomena such as "long distance movement". One may very well imagine that an optimal matching algorithm does not operate exclusively on constructions which are explicitly present in the surface structure of the sentence; "transformations" (in the classical Chomskyan sense) may very well play a role in the parsing process.

To demonstrate the technological usefulness of the matching algorithm, its disambiguating capacities must be compared with the capacities of existing methods: conventional grammars combined with nth order Markov-probabilities; and probabilistic contextfree grammars, which already have produced encouraging results (Fusijaki et al., 1989).

2.

Some questions about the status of these mechanisms in the context of the Theory of Formal Languages deserve explicit study. First of all it is not a priori obvious whether matching algorithms based on the ideas sketched above will define formal languages which (apart from the plausibility ordering between different structures) could in principle also be defined by existing grammatical formalisms, and, if so, by which ones. One can also wonder about stronger equivalence-relations: is it perhaps provable that applying a matching-algorithm to a corpus that was generated by a certain grammar, assigns the same analyses to input-sentences as the grammar?

Because the matching-algorithms define (partially) ordered sets of sentences (in a possibly non-deterministic way), one may think of extending the Theory of Formal Languages so as to include these aspects, which so far were considered to fall outside the realm of Mathematical Linguistics.

3.

The ideas sketched here are of eminent importance for psycholinguistics. Though Chomsky has always emphasized the abstract nature of his theory of language, psychologists have often interpreted the rule systems of linguistics as "psychologically real". Data-oriented parsing is psychologically much more plausible, because it does not work with abstract rules , but with concrete language experiences. This has interesting consequences, for instance, for language acquisition. Because we do not assume an abstract grammar, there is a complete continuity between the early and the later stages of language use. We do not postulate a separate process of language acquisition. This raises an important question, of course, about the nature of early language use: how does the matching algorithm work when there is no corpus yet? This question brings the non-linguistic component of language processing into the picture: the projection of the semantic/pragmatic context onto the linguistic input. In the early stages of language use this component is the dominant one -- later the linguistic component gets icreasingly prominent. (That is why adult language users can have "grammaticality judgments" about contextless "example sentences". Beginning language users would not be able to do this.)

Our perspective creates the possibility for a plausible model of language acquisition: the gradual development of the linguistic component of language processing, as a result of the gradual growth of the repertoire of linguistic experiences, and the increasing complexity of these experiences. But it will not be simple to describe in detail how language processing takes place in the early, pragmatically and semantically oriented stages, and how the later, more structure-oriented strategies, get bootstrapped out of that.

4.

The language processing techniques proposed here are probably much more "robust" than the

techniques which are currently used. They may be better capable of dealing with mistakes, i.e., they may be able to make a guess about the proper analysis of errorful input.

Robustness is practically important, because minor typos and ungrammaticalities often occur in interactive alphanumeric computer-input, and because the system's characterisation of the expected input (whether it be a grammar or a corpus) will always have undesired limitations. In processing spoken language these problems are even worse.

5.

This approach will ideally yield algorithms that produce provably optimal analyses of their input. But we do not get this result for free. It requires an amount of computation that may be hardly feasible with current equipment. The implementation of annotated corpora and matching-algorithms is therefore a major issue. First we will have to be concerned with efficient implementation on existing hardware. For really large-scale application of these techniques, the development of special-purpose hardware may be necessary.

9. References.

L.B. Bahl, F. Jelinek, and R.L. Mercer: "A maximum likelihood approach to continuous speech recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, No.2, March 1983.

A.L. Becker: "Biography of a sentence: a Burmese proverb." In: E.M. Bruner (ed.): *Text, play, and story: The construction and reconstruction of self and society*. Washington, D.C.: American Ethnology Society, 1984a. Pp. 135-155.

A.L. Becker: "The linguistics of particularity: Interpreting superordination in a Javanese text." *Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society*, pp. 425-436. Berkeley, Cal.: Linguistics Department, University of California at Berkeley, 1984b.

S. Boisen, Y. Chow, A. Haas, R. Ingria, S. Roukos, R. Scha, D. Stallard, en M. Vilain: *Integration of Speech and Natural Language. Final Report*. BBN Systems and Technologies Corporation, Cambridge, Mass. Report no. 6991, March 1989.

D. Bolinger: "Syntactic blends and other matters." *Language* **37**, 3 (1961), pp. 366-381.

D. Bolinger: "Meaning and Memory." *Forum Linguisticum* **1**, 1 (1976), pp. 1-14.

N. Chomsky: *Syntactic Structures*. The Hague: Mouton, 1957.

A.M. Derouault and B. Merialdo: "Natural language modeling for phoneme-to-text transcription." *IEEE Trans. PAMI*, 1986.

C.J. Fillmore, P. Kay, and M.C. O'Connor: "Regularity and idiomaticity in grammatical constructions." *Language*, **64**, 3 (1988)

J.A. Fodor: *The language of thought*. New York: T.Y. Crowell, 1975.

T. Fusijaki: "A stochastic approach to sentence parsing." *Proc. 10th International Conference on Computational Linguistics*. Stanford, CA, 1984.

T. Fusijaki, F. Jelinek, J. Cocke, E. Black, and T. Nishino: "A Probabilistic Parsing Method for Sentence Disambiguation." *Proc. International Parsing Workshop '89*. Pittsburgh: Carnegie-Mellon University, 1989. Pp. 85-94.

P. Hopper: "Emergent Grammar." *Proceedings of the 13th Annual Meeting of the Berkeley Linguistics Society*. Berkeley, Cal.: Linguistics Department, University of California at Berkeley, 1987.

F. Jameson: *The prison-house of language: A critical account of structuralism and Russian Formalism*. Princeton and Londen: Princeton University Press, 1972.

F. Jelinek: *Self-organized language modeling for speech recognition*. Ms., IBM T.J. Watson Research Center, Yorktown Heights, N.Y., 1986.

W.J.M. Levelt: "The connectionist fashion. Symbolic and subsymbolic models of human behaviour." In: C. Brown, P. Hagoort, and Th. Meijering (ed.): *Vensters op de geest. Cognitie op het snijvlak van filosofie en psychologie*. Utrecht: Stichting Grafiet, 1989. Pp. 202-219. [In Dutch]

R.J. Matthews: "Are the grammatical sentences of a language a recursive set?" *Synthese* **40** (1979), pp. 209-224.

J.L. McClelland, D.E. Rumelhart, and the PDP Research Group: *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*. Cambridge, Mass.: MIT Press, 1986.

S. Pinker en J. Mehler (red.): *Connections and Symbols*. Cambridge, Mass.: MIT Press, 1988.

F. Robert: "La langue." In: D. Causset et al.: *La Linguistique*. Parijs: Librairie Larousse, 1977.

D.E. Rumelhart, J.L. McClelland, and the PDP Research Group: *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, Mass.: MIT Press, 1986.

S.P. Stich: "What every speaker knows." *Philosophical Review*, 1971.