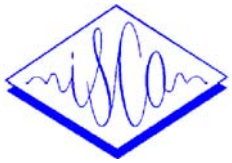

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

NPARSE -- A SHALLOW N-GRAM-BASED GRAMMATICAL-PHRASE PARSER

Alice Carlberger
Centre for Speech Technology
Department of Speech, Music and Hearing, KTH
SE-100 44 Stockholm, Sweden
alice@speech.kth.se
http://www.speech.kth.se

ABSTRACT

Nparse is a shallow probabilistic unification-based parser for N-best list resorting and the finding of simple grammatical phrases. It is data-driven and robust, allowing both domain-specific and unrestricted-language training. We believe it can be an interesting alternative for use in a synthesis or recognition front end. This parser has been trained for Swedish on a fine-grained set of grammatical-phrase nodes and grammatical features and evaluated on three language domains. A tree bank database has been built and a detailed linguistic assessment performed. Later, these results will be compared with evaluation on a simplified node-and-feature system. Our aim is to find the optimal system complexity for accurately establishing phrase boundaries and phrase types in newspaper text and, ultimately, unrestricted language. For this, a combination of iterative manual training and unsupervised training will be used.

Keywords: grammar parsing, NLP, Swedish

1. INTRODUCTION

The term "natural-language parsing" usually refers to the segmentation of text or speech into hierarchically arranged multi-layered syntactic chunks. The goal of most parsing is, thus, to provide a syntactic analysis, which is then used, e.g., in speech recognition and synthesis, dialogue, machine translation, and information retrieval systems. In contemporary literature, the concept of "parsing" exists under several guises. Furthermore, these terms are sometimes used to refer to the linguistic formalism, sometimes to the empirical methodology used. Examples of parsers for Swedish are those found in SVENSK [1], ChartParser [2], Fully Incremental Parsing [3], the Swedish Core Language Engine (SLE) [4], and Swedish Constraint Grammar (SWECG) [5].

Essential to parsing is the ability to distinguish between identical, i.e., homographic, word entries in the lexicon. In one study, approximately 645,000 out of 1,000,669 tokens in running Swedish newspaper text were shown to be homographic [6]. Homography can occur between lexicon entries (e.g., "sticka" *splinter; knitting needle* (noun) and "sticka" *to stick; to knit* (verb)), between inflections or between a lexicon entry and an inflection (e.g., "sticka" *to knit; knit* lexicon entry/infinitive versus imperative).

In Swedish, the noun and adjective are encoded for gender (neuter or common), number (singular or plural), and definiteness (definite or indefinite). Their inflectional paradigms typically have eight and six forms, respectively. Whereas the indefinite article is a separate word as in English, the definite article is most often encoded as a gender-number-sensitive nominal suffix. Verbs are encoded for tense (present, preterit, past participle, and supine) and mode (indicative and imperative); unlike many other Indo-European languages, they are not encoded for person and number. A typical verb paradigm consists of five inflections.

Major homographic areas are preposition versus verb particle (see next paragraph), adjective versus adverb, adjective or past participle versus its nominalization, article versus pronoun, infinitive versus imperative, and noun versus verb in the infinitive. Disambiguation is complicated by the frequent absence of distinguishing contextual information at the lexical, grammatical or syntactic level.

Like other Germanic languages, Swedish uses particle verbs as an integral part of its makeup. These are verbs with a postpositional particle, which in some forms is prefixed to the verb and essential to the concept of the verb. Frequently, the particle is a preposition homograph. Most of these verbs have nonparticled counterparts representing a distinct concept. A major issue is that the latter are often followed by a *true* preposition. For example, "Han *steg på* bussen" translates as *He got on the bus,* whereas "Han steg *på gången"* translates as *He stepped on the path.* In the former case, "på" is a particle, belonging to the verb phrase and crucial to the concept of the verb. In the latter, it is a preposition, thus belonging to the prepositional phrase. The distinction is semantic as well as prosodic.

In light of these facts, a parser can be an interesting alternative for use in a synthesis or recognition front end. For this purpose, a simple probabilistic parser, Nparse, was developed [7]. Our goal is not complex hierarchical syntactic analysis, but N-best list re-sorting and shallow segmentation into simple grammatical phrases. (Throughout the paper, "phrase" and "node" are used interchangeably. The term "sentence" denotes both sentences and utterances.) Nparse uses nodes and arcs to analyze the input and assigns structure also to those sentences for which it cannot offer a complete analysis. Since the functionality of recognition front ends is based on simple n-gram analysis, we believe that Nparse could be a suitable way to improve grammatical analysis. The system is data-driven and robust and allows both domain-specific and general-language training.

This paper describes the example-based training and testing of Nparse on Swedish, and the resulting tree bank database. The aim of our work is to determine the degree of language complexity the system can handle. We do this by successively decreasing the detail of grammatical analysis. In this paper, parse results from testing a fine-grained version on three language domains are presented, along with an analysis of the linguistic issues at hand. The first and most important domain is a combination of informal written and newspaper text for use in a text-to-speech application. The second domain is transcribed spontaneous speech collected in a dialogue system on public display, and the third is structured language for a speech recognition interface to an engineering design program.

2. THE PARSER

2.1 Functionality

Nparse accepts text or word graphs (from, e.g., a speech recognizer) as input and, using lookup in a feature-based lexicon, creates word graphs itself during processing. In this manner, homograph interpretations as well as multiple-word interpretations, such as lexicalized phrases and proper names, are represented in separate paths. The resulting possible sequences are assigned probabilities based on statistics obtained earlier in the training process.

Four levels of operation are used, namely, top, phrase, preterminal, and word. Phrase and preterminal names are defined using unification and a flexible set of features. The features, which can be grammatical, syntactic, morphological or semantic, are used to attach words to preterminals.

Nparse finds grammatical-phrase boundaries through n-gram analysis of unit sequences, where a unit is either a preterminal or a phrase node. It bases the calculation of the preterminal probability on the foregoing sequence of both preterminals and phrase nodes. The phrase node probability, on the other hand, is based only on the sequence of phrase nodes. Finally, the hypothesis probability is calculated as a smoothed sum of logarithmic probabilities. The functionality is illustrated in Figure 1, which shows a parse tree of the sentence "Hon tycker om små grodor" *(Lit. She thinks <particle> small toads,* i.e., *She likes small toads.)* Constituents of a node (i.e., phrase) are indicated by solid black lines; as seen, the verb phrase consists of verb and particle (or clause adverb, in our system; see below), and the object phrase, of adjective and noun.

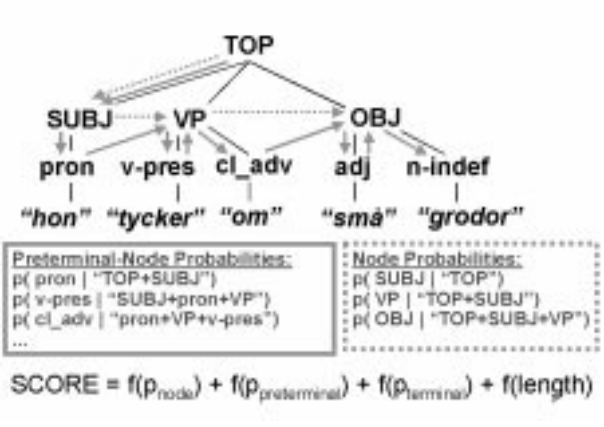

Figure 1. Parse Tree

2.2 Training

The probabilities are trained by a tree bank, which consists of n-grams that have been collected and represented in a tree structure. Retrieving a probability involves searching for the n-gram of maximum length in the n-gram tree. Words not in the lexicon are skipped, as are units incongruous with the n-gram statistics.

Parse training is an iterative process with manual correction of the training corpus after each run. As a base for this bootstrap procedure, a set of 460 n-gram sequences was formed manually. Three language domains were chosen out of convenience, and Nparse was trained and tested on these. See Section 3. The resulting tree bank database consists of approximately 3000 unique sentences and 9000 unique number-unit expressions. Included in the bootstrapping is also a continuous revision of the node-and-feature system and lexicon. The results from training and testing Nparse on the fine-grained node system are presented in Section 3.

2.3 Features, Nodes, and Unification for Swedish

In the fine-grained version of Nparse, 50 grammatical-morphological features and 28 nodes were used. Of the features, 25 were used to attach a word or group of words directly to the preterminal. Two of the nodes (OBJTOP and VPTOP) were based strictly on word order, overriding their grammatical counterparts (OBJECT and all types of verb phrases). This was necessary to reduce the variations in word order and consequent disruption of n-gram statistics. The node system is discussed in detail in Section 3.

Left-to-right, feature-based unification constrains the sequence of units at the preterminal and node levels or at both levels simultaneously. Unification handles, e.g., noun phrase-internal gender-number-definiteness agreement. It also constrains the appearance of subject complement versus object after a verb phrase. However, unification can be disrupted, mainly by certain types of main-clause-internal subclauses and noncanonical word order.

3. TRAINING AND TESTING ON THREE DOMAINS

3.1 Informal/News Text

Largest of the three domains is the informal/news text. Our main goal, here, has been to train Nparse for a text-to-speech system. To this end, we have used two different written texts: a selection of newspaper text and a list of 1,078 training sentences that were created manually in an effort to capture the basic structures of general, basically informal written language. Characteristic of this domain compared to the others, is the high rate of statements, noncanonical word order, and main-clause-internal subclauses. The training file consists of the handcrafted sentences along with a random selection of 361 newspaper sentences. For testing, 200 sentences from the remaining 2,000-sentence newspaper corpus were used.

3.2 Spontaneous Speech

The second domain is spontaneous speech. A set of 516 utterances was collected in the August dialog system [8,9,10] and transcribed for parse training. The August system features an animated talking agent that provides information to visitors at the Stockholm Cultural Center. August can give the user information either about Stockholm or the research conducted at the Centre for Speech Technology (CTT.) The domain is

characterized by a relatively high rate of questions and imperatives, object-verb-subject statements (as opposed to the canonical subject-verb-object order), and usage of the pronominal, most often impersonal, "det" *it, that,* (as in "Det regnar." *It is raining).*

3.3 Structured Language

The parser is also being trained for use in a voice-controlled engineering design system, ICAD, in the EU-funded ENABL Project [11]. In this system, it is used as an add-on to re-sort N-best hypotheses produced by the recognizer, which uses a bigram grammar. (The recognizer and other modules of the speech recognition interface are being developed at KTH.) The language in this domain is spoken but subject to well-defined application-specific constraints on the syntax and lexicon. A total of 9,161 sentences were used for parse training, including 8,946 number-unit sentences. Numbers are handled by a system of 14 features and 60 basic lexicon entries, from which all other numbers are generated, including decimal numbers and fractions. Testing was conducted on a set of 415 sentences.

3.4 Results & Linguistic Analysis

Test data for the three domains are displayed in Table 1. The number of grammatically incorrectly parsed sentences (or utterances) is shown on line three. They are 101, 21, and 1, respectively, or 50.50%, 7.00%, and 0.24%, respectively, of the test sentences. The much higher rate for news text is undoubtedly due to its relatively complex sentence structure, which is also captured by the higher average number of words per test sentence: 8.31, compared with 4.25, and 4.40, for the other domains (line two). The focus of this study is on the news text, and data from the other two domains are used only for comparison. After parsing of the test text in the news domain, the grammatical error rate is 0.81 per sentence or 1.61 per incorrectly parsed sentence. Grammatical errors, then, are phrase label or phrase boundary errors, and can arise from the failure to disambiguate one or several words at the lexical level. A grammatical error can also be the result of misrepresentative n-gram statistics. A disambiguation error at the lexical level almost invariably results in a grammatical error. Skipping or misinterpretation of words not in the lexicon is, for our purposes, not considered an error, even if it leads to a grammatical error. Of the 163 grammatical errors, 116 are phrase boundary errors and 47 phrase label errors. The corresponding numbers for the other two domains are much smaller.

	In/News Text	**Spont Speech**	**Struct Lang**
Number of Test Snts	200	300	415
Avg Words per Snts	8.31	4.25	4.40
Number of Gram Incorr Snts	101	21	1
Number of Gram Errors	163	29	1
Avg Gram Error per Snts	0.81	0.10	0.002
Number of Phrase Boundary Errors	116	13	1
Number of Phrase Label Errors	47	16	0

Table 1. Results of Testing Nparse on Informal/News Text, Spontaneous Speech, and Structured Language

A breakdown of grammatical parse errors in the news text is shown in Table 2. Within each of the two phrase error categories (boundary errors and label errors), bold type indicates major categories, and regular type, subcategories. The first number column shows the actual number of errors, the second number column the percentage of each error type with respect to total grammatical error. As can be seen, boundary errors (71.17%) heavily outweigh label errors (28.83%). Here follow an analysis and discussion of each of their categories. It should be noted that these linguistic phenomena exhibit a complex interplay, which complicates classification.

		#	**% of Gr Er**
1. Phrase Boundary Errors		**116**	**71.17**
Incorrect Splitting of NP's		***35***	***21.47***
due to tagging **art** as **pron**		15	9.20
due to tagging **adj** as **noun**		8	4.91
due to tagging **other** as **noun**		12	7.36
Preposition - Verb Particle		***33***	***20.24***
Adjective or Adverb?		***13***	***7.98***
	adj - adv	7	4.29
	adv - adj	6	3.68
Adjective - Verb		***6***	***3.68***
Numbers		***5***	***3.07***
Adverb after Verb		***4***	***2.45***
Other		***20***	***12.28***
2. Phrase Label Errors		**47**	**28.83**
Mislabeling of NP's		***14***	***8.59***
	obj - subj	10	6.13
	other	4	2.45
Nonnoun - Noun		***11***	***6.75***
Noun - Nonnoun		***7***	***4.29***
Other		***15***	***9.20***

Table 2. Grammatical Error Distribution in Test on Informal/News Text

3.4.1 Phrase Boundary Errors

One of the two largest categories of boundary errors is the incorrect splitting of a noun phrase (21.47%). This arises mainly when an article is misinterpreted as a pronoun or an adjective as a nominalization. In these cases, a contributing factor is the homography of verb participles with adjectives and their nominalizations. An example is "den jagade katten", which has two alternate readings: as a noun phrase consisting of an article, past participle (adjective) and noun *(the chased cat,* i.e., *the cat that was or had been chased)* and as a sentence consisting of pronoun, preterit verb, and noun *(It chased the cat).*

The other large boundary error category is the misinterpretation of a preposition as a verb particle (classified as a clause adverb in our system) that is therefore incorporated into the end of a verb phrase instead of into the beginning of a prepositional phrase (20.24%). This is, in turn, because our system distinguishes between two types of adverbs: those whose scope invariably is the whole preceding clause, i.e., clause adverbs (e.g., "dock" *however)* and those that can have either a following noun phrase or the whole clause as their scope (e.g., "snabbt" *fast, rapidly).* As in English, the first group is very small, while the second one is infinitely large, due to creation

by means of any of three adjectival inflections: positive neuter singular indefinite ("snabbt" *fast, rapidly),* comparative ("snabbare" *faster)* or superlative ("snabbast" *fastest*). In light of this, the particle-versus-preposition issue can be said to overlap with a closely related issue, which, however, was not in evidence in the test results. It is the issue of distinguishing whether an adverb belongs to the preceding verb phrase or the following noun phrase.

Directly related to the productive adverb formation process is the third major error category. It is the failure to distinguish between the adjectival versus adverbial function (7.98%). The next error category is due to the misinterpretation of an adjective as a verb (3.68%), as in "hela hösten", which was interpreted by Nparse as **heal the autumn* ("hela" = verb) instead of as *the whole autumn* ("hela" = adjective). Furthermore, certain phrases containing number units have been incorrectly split due to the fact that the number grammar has not yet been incorporated into this domain (3.07%). Finally, boundary errors arise because of misanalysis of a postverbal clause adverb as a separate adverbial phrase rather than as part of the verb phrase (2.45%).

3.4.2 Phrase Label Errors

Phrase label errors are mainly the mislabeling of noun phrases (8.59%) and the misinterpretation of a noun phrase as a non-noun phrase, specifically an adverbial, verb, or adjective phrase and vice versa (6.75% and 4.29%, respectively). Noun phrases are of three kinds: subject, object, and subject complement. Noun phrase mislabeling, then, most often entails the misreading of an object as a subject. This is undoubtedly due to the high incidence of object fronting in the test text compared with training text.

4. CONCLUSIONS & FUTURE WORK

With the fine-grained node system presented here, it seems that Nparse could work well for structured domains such as engineering design. However, during training and testing on more unrestricted text, it has become clear that trying to improve coverage of certain structures may result in overgeneration for some texts or reduced coverage of others. Since the ultimate goal is to reach a level of training that represents the general language, either larger amounts of training data or a less complex node system, or a combination of both, is needed. As indicated by the test results, the major low-resolution areas that need to be tended to are nominalization, verb particles, and adverbialization.

As for noun phrase labeling, replacing SUBJECT, OBJECT, and OBJTOP with NP (noun phrase) would not only reduce their confusion with each other but also avoid the necessity of distinguishing them from nominal adverbial phrases such as "en vecka senare" *a week later*. Regarding verb particle-preposition homographs, they could constitute their own pre-terminal class, instead of being treated either as clause adverbs or prepositions. This would not eliminate, but at least restrict, the ambiguity problem.

ACKNOWLEDGEMENTS

Nparse has been designed and developed by Rolf Carlson. I am grateful for his guidance and constructive feedback.

REFERENCES

[1] Eriksson, M. and Gambäck, B. (1997) SVENSK: A Toolbox of Swedish Language Processing Resources. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing,* Tzigov Chark, Bulgaria, September.

[2] Wirén, M. (1992) Studies in Incremental Natural-Language Analysis, Doctor of Philosophy Thesis, Linköping University, Dept. of Computer and Information Science, Linköping, Sweden, December.

[3] Wirén, M. (1994). Minimal Change and Bounded Incremental Parsing. In *Proceedings of the 15th International Conference on Computational Linguistics,* Kyoto, Japan: Vol. 1, pp. 461--467.

[4] Gambäck, B. and Rayner, M. (1992) The Swedish Core Language Engine. In Ahrenberg, L. (editor) *Papers from the 3rd Nordic Conference on Text Comprehension in Man and Machine,* Linköping University, Linköping, Sweden, April.

[5] Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995) Constraint Grammar: A language-independent system for parsing unrestricted text, Mouton de Gruyter, Berlin and New York.

[6] Allén, S. (editor) (1970) Frequency Dictionary of Present-Day Swedish, Almqvist & Wiksell, Stockholm, Sweden.

[7] Carlson, R. (1997), personal communication.

[8] Gustafson, J., Lundeberg, M., and Liljencrants, J. (1999), Experiences from the development of August -- A multimodal spoken dialogue system. To be published in *Proceedings of IDS'99, ESCA workshop on Interactive Dialogue in Multi-Modal Systems.*

[9] Bell, L. and Gustafson, J. (1999), Interacting with an animated agent: User strategies in a spoken dialogue system. To be published in *Proceedings of Eurospeech 99.*

[10] Gustafson, J., Lindberg, N., Lundeberg, M., Svensson, E.-L., and Öhman, T. (1999), The August Spoken Dialogue System. To be published in *Proceedings of Swedish phonetics conference Fonetik 99.*

[11] Carlberger, A. (1998), Grammar and Lexicons for a Speech-Interfaced Knowledge-Based Engineering Program (ICAD). In Loncke, F. T., Clibbens, J., Arvidson, H. H., Lloyd, L. L. (Eds.) (1999). *Augmentative and Alternative Communication: New Directions in Research and Practice.* London: Whurr Publishers.