EACL-2006

11th Conference
of the European Chapter of the
Association for Computational Linguistics

Proceedings of the workshop on

ROMAND 2006: Robust Methods in Analysis of Natural language Data

April, 3rd , 2006
Trento, Italy

The conference, the workshop and the tutorials are sponsored by:

Center for the Evaluation of Language and Communication Technologies

Celct
c/o BIC, Via dei Solteri, 38
38100 Trento, Italy
http://www.celct.it

Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France
http://www.xrce.xerox.com

CELI s.r.l.
Corso Moncalieri, 21
10131 Torino, Italy
http://www.celi.it

Thales
45 rue de Villiers
92526 Neuilly-sur-Seine Cedex, France
http://www.thalesgroup.com

EACL-2006 is supported by

Trentino S.p.a.  and Metalsistem Group 

Order copies of ACL proceedings from:
Priscilla Rasmussen,
Association for Computational Linguistics (ACL),
3 Landmark Center,
East Stroudsburg, PA 18301 USA

Phone +1-570-476-8006
Fax +1-570-476-0860
E-mail: acl@aclweb.org
On-line order form: http://www.aclweb.org/

Preface

Robustness is a fuzzy notion, which accordingly is difficult to define. This difficulty mainly arises from the fact that robustness touches upon the highly subjective and application-specific notion of the norm and the deviation thereof. Thus, robustness is inherently about the unexpected, about all the things that can and will go wrong, which have not been taken care of and which usually cannot be fully anticipated.

On the other hand, robustness is one of the most prominent characteristics of intelligent human behaviour which facilitates flexible and sensible responses to a wide variety of unpredictable situations. A closer look at the phenomenon reveals a multitude of different aspects. Thus merely speaking about robustness requires to identify precisely against which particular kind of deviation robust behaviour is desired, a question which is highly dependent on the application and the task at hand.

The contributions to this workshop deal with issues of robustness in quite different areas of Natural Language Processing, ranging from anaphora resolution on the one hand to three different sentence analysis tasks on the other.

Delmonte et al. compare their system for anaphora resolution against three other systems from the literature and show that it outperforms the other approaches significantly. They attribute this success to the use of a robust parser, which is able to determine surface and deep syntactic relationships robustly.

Semantically annotated structures are in the focus of Musillo and Merlo's paper. They modified a statistical parsing model to also assign semantic role labels as annotated in the Prop bank. The solution differs from other approaches in that it integrates both labeling tasks into a single processing step. The results show that despite a 20-fold increase in non-terminals a fairly high f-measure of 82\% was obtained. This corresponds to an absolute reduction of as little as 6\% compared to the baseline system, which only considers purely syntactic categories. Musillo and Merlo interpret this as evidence for the robustness of the underlying stochastic model (Simple Synchrony Network), which does not require making specific assumptions about parameter independence.

Philippe Blache applies Property Grammar, a constraint-based formalism for phrase structure descriptions, to shallow parsing of French sentences. Robustness in this case is achieved by relaxing constraints if necessary.

Finally, Foth and Menzel investigate the relationship between coverage and accuracy when parsing unrestricted German text. Their results confirm that even for a grammar which is able to determine the optimal structure according to some given criterion, there is a reciprocal correspondence: reducing the coverage by removing rare phenomena from the grammar slightly increased the accuracy of the parser. They claim that this finding provides support to the hypothesis that robustness, which in this case in introduced by means of weighted constraints, might be a more desirable property than coverage as long as really rare phenomena are considered.

With this selection the workshop unites samples of different techniques for achieving robustness for a range of different processing task. This, however, leaves completely

untouched the problem of measuring robustness properties as such. If robustness is defined as a smooth degradation in the performance of a system when faced with unexpected input, common evaluation procedures where test and training data are obtained from the same source, do not contribute very much to a deeper understanding of what robustness really means and how it can be achieved best.

Talking about robustness as the ability to deal with deviation from the norm naturally includes issues like scalability and portability. It therefore remains a challenge for future research to develop proposals for standardized scenarios in which such properties can be evaluated across a wide variety of languages and processing tasks. In this sense this workshop is a small contribution of an ongoing effort towards a common research goal, which step by step might become less elusive: How to make natural language processing systems more stable, more dependable, more useful, ...in short, more human like.

Wolfgang Menzel

Workshop Organizers:

Wolfgang Menzel, University of Hamburg, Germany
Vincenzo Pallotta, University of Fribourg, Switzerland

Programme Committee:

Afzal Ballim
Alberto Lavelli
Alexander Clark
Amedeo Cappelli
Amalia Todirascu
Atro Voutilainen
Beth-Ann Hockey
Bangalore Srinivas
Dan Cristea
Dan Tufis
Diego Mollá-Aliod
Eric Wehrli
Fabio Massimo Zanzotto
Fabio Rinaldi
Florentina Hristea
Frank Keller
Frank Schilder
Jean-Cédric Chappelier
Jean-Pierre Chanod
Joachim Niehren
John Dowding
Josè Iria
Kay-Uwe Carstensen
Maria Teresa Pazienza
Manny Rayner
Martin Kay
Michael Hess
Geertjan van Noord
Gian Lorenzo Thione
Günther Görz
Roberto Basili
Rodolfo Delmonte
Salah Ait-Mokhtar
Vincenzo Pallotta
Violeta Seretan
Wolfgang Menzel
Yuji Matsumoto

Invited speaker:

Gertjan van Noord, University of Groningen, The Netherlands

Endorsed by:

CELCT Center for the Evaluation of Language and Communication Technology.
http://www.celct.it/

Workshop website:

http://www.icsi.berkeley.edu/~vincenzo/romand2006/

Workshop Program

Monday, April 3rd

14:00-14:05 ***Welcome***

14:05-15:05 Invited Talk: ***Robust Parsing, Error Mining, Automated Lexical Acquisition, and Evaluation***, Gertjan van Noord

15:05-15:35 ***Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach***. Rodolfo Delmonte, Antonella Bristot, Marco Aldo Piccolino Boniforti, Sara Tonelli.

15:35-16:00 BREAK

16:00-16:30 ***Robust Parsing of the Proposition Bank***. Gabriele Musillo and Paola Merlo.

16:30- 17:00 ***A Robust and Efficient Parser for Non-Canonical Inputs.*** Philippe Blache

17:00-17:30 ***Robust Parsing: More with Less***. Kilian Foth and Wolfgang Menzel

17:30-18:30 ***Final Panel***

Table of Contents

Robust Parsing, Error Mining, Automated Lexical Acquisition, and Evaluation

Gertjan van Noord
University of Groningen
vannoord@let.rug.nl

Abstract

In our attempts to construct a wide coverage HPSG parser for Dutch, techniques to improve the overall robustness of the parser are required at various steps in the parsing process. Straightforward but important aspects include the treatment of unknown words, and the treatment of input for which no full parse is available.

Another important means to improve the parser's performance on unexpected input is the ability to learn from your errors. In our methodology we apply the parser to large quantities of text (preferably from different types of corpora), and we then apply error mining techniques to identify potential errors, and furthermore we apply machine learning techniques to correct some of those errors (semi-)automatically, in particular those errors that are due to missing or incomplete lexical entries.

Evaluating the robustness of a parser is notoriously hard. We argue against coverage as a meaningful evaluation metric. More generally, we argue against evaluation metrics that do not take into account accuracy. We propose to use variance of accuracy across sentences (and more generally across corpora) as a measure for robustness.

Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach

Rodolfo Delmonte, Antonella Bristot, Marco
Aldo Piccolino Boniforti, Sara Tonelli
Department of Language Sciences
Università Ca' Foscari – Ca' Bembo
30120, Venezia, Italy
delmont@unive.it

Abstract

In this paper we will present an evaluation of current state-of-the-art algorithms for Anaphora Resolution based on a segment of Susanne corpus (itself a portion of Brown Corpus), a much more comparable text type to what is usually required at an international level for such application domains as Question/Answering, Information Extraction, Text Understanding, Language Learning. The portion of text chosen has an adequate size which lends itself to significant statistical measurements: it is portion A, counting 35,000 tokens and some 1000 third person pronominal expressions. The algorithms will then be compared to our system, GETARUNS, which incorporates an AR algorithm at the end of a pipeline of interconnected modules that instantiate standard architectures for NLP. F-measure values reached by our system are significantly higher (75%) than the other ones.

1 Introduction

The problem of anaphora resolution (hence AR) looms more and more as a prominent one in unrestricted text processing due to the need to recover semantically consistent information in most current NLP applications. This problem does not lend itself easily to a statistical approach so that rule-based approaches seem the only viable solution.

We present a new evaluation of three state-of-the-art algorithms for anaphora resolution – GuiTAR, JavaRAP, MARS – on the basis of a portion of Susan Corpus (derived from Brown Corpus) a much richer testbed than the ones previously used for evaluation, and in any case a much more comparable source with such texts as newspaper articles and stories. Texts used previously ranged from scientific manuals to descriptive scientific texts and were generally poor on pronouns and rich on nominal descriptions. Two of the algorithms – GuiTAR and JavaRAP - use Charniak's parser output, which contributes to the homogeneity of the type of knowledge passed to the resolution procedure. MARS, on the contrary, uses a more sophisticated input, the one provided by Connexor FDG-parser. The algorithms will then be compared to our system, GETARUNS, which incorporated an AR algorithm at the end of a pipeline of interconnected modules that instantiate standard architectures for NLP. The version of the algorithm presented here is a newly elaborated one, and is devoted to unrestricted text processing. It is an upgraded version from the one discussed in Delmonte (1999;2002a;2002b) and tries to incorporate as much as possible of the more sophisticated version implemented in the complete GETARUN (see Delmonte 1990;1991;1992;1994; 2003;2004).

The paper is organized as follows: in section 2 below we briefly discuss architectures and criteria for AR of the three algorithms evaluated. In section 3 we present our system. Section 4 is dedicated to a compared evaluation and a general discussion.

2 The Anaphora Resolution Algorithms

We start by presenting a brief overview of three state-of-the-art algorithms for anaphora resolution – GuiTAR, JavaRAP, MARS.

2.1 JavaRAP

As reported by the authors (Long Qiu, Min-Yen Kan, Tat-Seng Chua, 2004) of the JAVA implementation, head-dependent relations required by RAP are provided by looking into the structural "argument domain" for arguments and into the structural "adjunct domain" for adjuncts. Domain information is important to establish disjunction relations, i.e. to tell whether a third person pronoun can look for antecedents within a certain structural domain or not. According to Binding Principles, Anaphors (i.e. reciprocal and reflexive pronouns),

must be bound – search for their binder-antecedent – in their same binding domain – roughly corresponding to the notion of structural "argument/adjunct domain". Within the same domains, Pronouns must be free. Head-argument or head-adjunct relation is determined whenever two or more NPs are sibling of the same VP.
Additional information is related to agreement features, which in the case of pronominal expressions are directly derived. As for nominal expressions, features are expressed in case they are either available on the verb – for SUBJect NPs– or else if they are expressed on the noun and some other tricks are performed for conjoined nouns. Gender is looked up in the list of names available on the web. This list is also used to provide the semantic feature of animacy.
RAP is also used to find pleonastic pronouns, i.e. pronouns which have no referents. To detect conditions for pleonastic pronouns a list of patterns is indicated, which used both lexical and structural information.
Salience weight is produced for each candidate antecedent from a set of salience factors. These factors include main Grammatical Relations, Headedness, non Adverbiality, belonging to the same sentence. The information is computed again by RAP, directly on the syntactic structure. The weight computed for each noun phrase is divided by two in case the distance from the current sentence increases. Only NPs contained within a distance of three sentences preceding the anaphor are considered by JavaRAP.

2.2 GuiTAR

The authors (Poesio, M. and Mijail A. Kabadjov 2004) present their algorithm as an attempt at providing a domain independent anaphora resolution module, "that developers of NLE applications can pick off the shelf in the way of tokenizers, POS taggers, parsers, or Named Entity classifiers". For these reasons, GuiTAR has been designed to be as independent as possible from other modules, and to be as modular as possible, thus "allowing for the possibility of replacing specific components (e.g., the pronoun resolution component)".
The authors have also made an attempt at specifying what they call the Minimal Anaphoric Syntax (MAS) and have devised a markup language based on GNOME mark-up scheme. In MAS, Nominal Expressions constitute the main processing units, and are identified with the tag NE <ne>, which have a CAT attribute, specifying the NP type: the-np, pronoun etc., as well as Person, Number and Gender attributes for agreement features. Also the internal structure of the NP is marked with Mod and NPHead tags.
The pre-processing phase uses a syntactic guesser which is a chunker of NPs based on heuristics. All NEs add up to a discourse model – or better History List - which is then used as the basic domain where Discourse Segments are contained. Each Discourse Segment in turn may be constituted by one or more Utterances. Each Utterance in turn contains a list of forward looking centers Cfs.
The Anaphora Resolution algorithm implemented is the one proposed by MARS which will be commented below. The authors also implemented a simple algorithm for resolving Definite Descriptions on the basis of the History List by a same head matching approach.

2.3 MARS

The approach is presented as a knowledge poor anaphora resolution algorithm (Mitkov R. [1995;1998]), which makes use of POS and NP chunking, it tries to individuate pleonastic "it" occurrences, and assigns animacy. The weighting algorithm seems to contain the most original approach. It is organized with a filtering approach by a series of indicators that are used to boost or reduce the score for antecedenthood to a given NP. The indicators are the following ones:
FNP (First NP); INDEF (Indefinite NP); IV (Indicating Verbs); REI (Lexical Reiteration); SH (Section Heading Preference); CM (Collocation Match); PNP (Prepositional Noun Phrases); IR (Immediate Reference); SI (Sequential Instructions); RD (Referential Distance); TP (Term Preference), As the author comments, antecedent indicators (preferences) play a decisive role in tracking down the antecedent from a set of possible candidates. Candidates are assigned a score (-1, 0, 1 or 2) for each indicator; the candidate with the highest aggregate score is proposed as the antecedent.
The authors comment is that antecedent indicators have been identified empirically and are related to salience (definiteness, givenness, indicating verbs, lexical reiteration, section heading preference, "non- prepositional" noun phrases), to structural matches (collocation, immediate reference), to referential distance or to preference of terms. However it is clear that most of the indicators have been suggested for lack of better information, in particular no syntactic constituency was available.
In a more recent paper (Mitkov et al., 2003) MARS has been fully reimplemented and the indicators updated. The authors seem to acknowledge the fact that anaphora resolution is a much more difficult task than previous work had suggested, In

unrestricted text analysis, the tasks involved in the anaphora resolution process contribute a lot of uncertainty and errors that may be the cause for low performance measures.
The actual algorithm uses the output of Connexor's FDG Parser, filters instances of "it" and eliminates pleonastic cases, then produces a list of potential antecedents by extracting nominal and pronominal heads from NPs preceding the pronoun. Constraints are then applied to this list in order to produce the "set of competing candidates" to be considered further, i.e. those candidates that agree in number and gender with the pronoun, and also obey syntactic constraints. They also introduced the use of Genetic Algorithms in the evaluation phase.
The new version of MARS includes three new indicators which seem more general and applicable to any text, so we shall comment on them.
Frequent Candidates (FC) – this is a boosting score for most frequent three NPs; Syntactic Parallelism (SP) – this is a boosting score for NPs with the same syntactic role as the pronoun, roles provided by the FDG-Parser; Boost Pronoun (BP) – pronoun candidates are given a bonus (no indication of conditions for such a bonus).
The authors also reimplemented in a significant way the indicator First NPs which has been renamed, "Obliqueness (OBL) – score grammatical functions, SUBJect > OBJect > IndirectOBJect > Undefined".
MARS has a procedure for automatically identifying pleonastic pronouns: the classification is done by means of 35 features organized into 6 types and are expressed by a mixture of lexical and grammatical heuristics. The output should be a fine-grained characterization of the phenomenon of the use of pleonastic pronouns which includes, among others, discourse anaphora, clause level anaphora and idiomatic cases.
In the same paper, the authors deal with two more important topics: syntactic constraints and animacy identification.

3 GETARUNS

In a number of papers (Delmonte 1990;1991; 1992;1994; 2003;2004) and in a book (Delmonte 1992) we described our algorithms and the theoretical background which inspired it. Whereas the old version of the system had a limited vocabulary and was intended to work only in limited domains with high precision, the current version of the system has been created to cope with unrestricted text. In Delmonte (2002), we reported preliminary results obtained on a corpus of anaphorically annotated texts made available by R.Mitkov on his website. Both definite descriptions and pronominal expressions were considered, success rate was at 75% F-measure. In those case we used a very shallow and robust parser which produced only NP chunks which were then used to fire anaphoric processes. However the texts making up the corpus were technical manuals, where the scope and usage of pronominal expressions is very limited.
The current algorithm for anaphora resolution works on the output of a complete deep robust parser which builds an indexed linear list of dependency structures where clause boundaries are clearly indicated; differently from Connexor, our system elaborates both grammatical relations and semantic roles information for arguments and adjuncts. Semantic roles are very important in the weighting procedures. Our system also produces implicit grammatical relations which are either controlled SUBJects of untensed clauses, arguments or adjuncts of relative clauses.
As to the anaphoric resolution algorithm, it is based on the original Sidner's (1983:Chapter 5) and Webber's (1983:Chapter 6) intuitions on Focussing in Discourse. We find distributed, local approaches to anaphora resolution more efficient than monolithic, global ones. In particular we believe that due to the relevance of structural constraints in the treatment of locally restricted classes of pronominal expressions, it is more appropriate to activate different procedures which by dealing separately with non-locally restricted classes also afford separate evaluation procedures. There are also at least two principled reasons for the separation into two classes.
The first reason is a theoretical one. Linguistic theory has long since established without any doubt the existence in most languages of the world of at least two classes: the class of pronouns which must be bound locally in a given domain and the class of pronouns which must be left free in the same domain – as a matter of fact, English also has a third class of pronominals, the so-called long-distance subject-of-consciousness bound pronouns (see Zribi-Hertz A., 1989);
The second reason is empirical. Anaphora resolution is usually carried out by searching antecedents backward w.r.t. the position of the current anaphoric expression. In our approach, we proceed in a clause by clause fashion, weighting each candidate antecedent w.r.t. that domain, trying to resolve it locally. Weighting criteria are amenable on the one hand to linear precedence constraints, with scores assigned on a functional/semantic basis. On the other hand, these criteria may be overrun by a functional ranking of clauses which requires to treat main clauses differently from secondary clauses,

and these two differently from complement clauses. On the contrary, global algorithms neglect altogether such requirements: they weight each referring expression w.r.t. the utterance, linear precedence is only physically evaluated, no functional correction is introduced.

3.1 Referential Policies and Algorithms

There are also two general referential policy assumption that we adopt in our approach: The first one is related to pronominal expressions, the second one to referring expressions or entities to be asserted in the History List, and are expressed as follows:

- no more than two pronominal expressions are allowed to refer back in the previous discourse portion;
- at discourse level, referring expressions are stored in a push-down stack according to Persistence principles.

Persistence principles respond to psychological principles and limit the topicality space available to user w.r.t. a given text. It has a bidimensional nature: it is determined both in relation to an overall topicality frequency value and to an utterance number proximity value.

Only "persistent" referring expressions are allowed to build up the History List, where persistence is established on the basis of the frequency of topicality for each referring expression which must be higher than 1. All referring expression asserted as Topic (Secondary, Potential) only once are discarded in case they appeared at a distance measured in 5 previous utterances. Proximate referring expressions are allowed to be asserted in the History List.

In particular, if Mitkov considers the paragraph as the discourse unit most suitable for coreferring and cospecifying operation at discourse level, we prefer to adopt a parameterized procedure which is definable by the user and activated automatically: it can be fired within a number that can vary from every 10 up to 50 sentences. Our procedure has the task to prune the topicality space and reduce the number of perspective topic for Main and Secondary Topic. Thus we garbage-collect all non-relevant entities. This responds to the empirically validated fact that as the distance between first and second mention of the same referring expression increases, people are obliged to repeat the same linguistic description, using a definite expression or a bare NP. Indefinites are unallowed and may only serve as first mention; they can also be used as bridging expression within opaque propositions. The first procedure is organized as follows:

A. For each clause,

1. we collect all referential expressions and weight them (see B below for criteria) – this is followed by an automatic ranking;
2. then we subtract pronominal expressions;
3. at clause level, we try to bind personal and possessive pronouns obeying specific structural properties; we also bind reflexive pronouns and reciprocals if any, which must be bound obligatorily in this domain;
4. when binding a pronoun, we check for disjointness w.r.t. a previously bound pronoun if any;
5. all unbound pronouns and all remaining personal pronouns are asserted as "externals", and are passed up to the higher clause levels;

B. Weighting is carried out by taking into account the following linguistic properties associated to each referring expression:

1. Grammatical Function with usual hierarchy (SUBJ > ARG_MOD > OBJ > OBJ2 > IOBJ > NCMOD);
2. Semantic Roles, as they have been labelled in FrameNet, and in our manually produced frequency lexicon of English;
3. Animacy: we use 75 semantic features derived from WordNet, and reward Human and Institution/Company labelled referring expressions;
4. Functional Clause Type is further used to introduce penalties associated to those referring expressions which don't belong to main clause.

C. Then we turn at the higher level – if any -, and we proceed as in A., in addition

1. we try to bind pronouns passed up by the lower clause levels
   - if successful, this will activate a retract of the "external" label and a label of "antecedenthood" for the current pronoun with a given antecedent;
   - the best antecedent is chosen by recursively trying to match features of the pronoun with the first available antecedent previously ranked by weighting;
   - here again whenever a pronoun is bound we check for disjointness at utterance level.

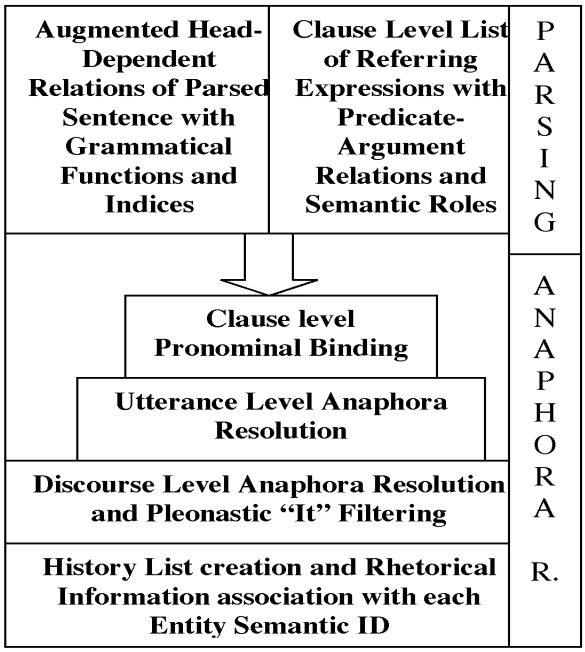
D. This is repeated until all clauses are examined and all pronouns are scrutinised and bound or left free.

E. Pronouns left free – those asserted as externals – will be matched tentatively with the best candidates provided this time by a "centering-like" algorithm.

Step A. is identical and is recursively repeated until all clauses are processed.

Then, we move to step B. which in this case will use all referring expressions present in the utterance, rather than only those available locally.

Fig. 1 GETARUNS AR algorithm



3.2 Focussing Revisited

Our version of the focussing algorithm follows Sidner's proposal (Sidner C., 1983; Grosz B., Sidner C., 1986), to use a Focus Stack, a certain Focus Algorithm with Focus movements and data structures to allow for processing simple inferential relations between different linguistic descriptions co-specifying or coreferring to a given entity.

Our Focus Algorithm is organized as follows: for each utterance, we assert three "centers" that we call Main, Secondary and the first Potential Topic, which represent the best three referring expressions as they have been weighted in the candidate list used for pronominal binding; then we also keep a list of Potential Topics for the remaining best candidates. These three best candidates repositories are renovated at each new utterance, and are used both to resolve pronominal and nominal cospecification and coreference: this is done both in case of strict identity of linguistic description and of non-identity. The second case may occur either when derivational morphological properties allow the two referring expressions to be matched successfully, or when a simple hyponym/hypernym relation is entertained by two terms, one of which is contained in the list of referring expressions collected from the current sentence, and the other is among one of the entities stored in the focus list.

The Main Topic may be regarded the Forward Looking Center in the centering terminology or the Current Focus. All entities are stored in the History List (HL) which is a stack containing their morphological and semantic features: this is not to be confused with a Discourse Model - what we did in the deep complete system anaphora resolution module – which is a highly semantically wrought elaboration of the current text. In the HL every new entity is assigned a semantic index which identifies it uniquely. To allow for Persistence evaluation, we also assert rhetorical properties associated to each entity, i.e. we store the information of topicality (i.e. whether it has been evaluated as Main, Secondary or Potential Topic), together with the semantic ID and the number of the current utterance. This is subsequently used to measure the degree of Persistence in the overall text of a given entity, as explained below.

In order to decide which entity has to become Main, Secondary or Potential Topic we proceed as follows:

- we collect all entities present in the History List with their semantic identifier and feature list and proceed to an additional weighting procedure;
- nominal expressions, they are divided up into four semantic types: definite, indefinite, bare NPs, quantified NPs. Both definite and indefinite NP may be computed as new or old entity according to contextual conditions as will be discussed below and are given a rewarding score;
- we enumerate for each entity its persistence in the previous text, and keep entities which have frequency higher than 1, we discard the others;
- we recover entities which have been asserted in the HL in proximity to the current utterance, up to four utterances back;
- we use this list to "resolve" referring expressions contained in the current utterance;
- if this succeeds, we use the "resolved" entities as new Main, Secondary, and Potential Topics and assert the rest in the Potential Topics stack;
- if this fails – also partially – we use the best candidates in the weighted list of referring expressions to assert the new Topics. It may be the case that both resolved and current best candidates are used, and this is by far the most common case.

4. Evaluation and General Discussion

Evaluating anaphora resolution systems calls for a reformulation of the usual parameters of Precision and Recall as introduced in IR/IE field: in that case, there are two levels that are used as valuable results; a first stage where systems are measured for their

capacity to retrieve/extract relevant items from the corpus/web (coverage-recall). Then a second stage follows in which systems are evaluated for their capacity to match the content of the query (accuracy-precision). In the field of IR/IE items to be matched are usually constituted by words/phrases and pattern-matching procedures are the norm. However, for AR systems this is not sufficient and NLP heavy techniques are used to get valuable results. As Mitkov also notes, this phase jeopardizes the capacity of AR systems to reach satisfactory accuracy scores simply because of its intrinsic weakness: none of the off-the-shelf parsers currently available overcomes 90% accuracy.

To clarify these issues, we present here below two Tables: in the first one we report data related to the vexed question of whether pleonastic "it" should be regarded as part of the task of anaphora resolution or rather part of a separate classification task – as suggested in a number of papers by Mitkov. In the former case, they should contribute to the overall anaphora resolution evaluation metrics; in the latter case they should be compute separately as a case of classification over all occurrences of "it" in the current dataset and discarded from the overall count. Even though we don't agree fully with Mitkov's position, we find it useful to deal with "it" separate, due to its high inherent ambiguity. Besides, it is true that the AR task is not like any Information Retrieval task.

In Table 1 below we reported figures for "it" in order to evaluate the three algorithms in relation to the classification task. Then in Table 2. we report general data where we computed the two types of accuracy reported in the literature. In Table 1 we split results for "it" into Wrong Reference vs Wrong Classification: following Mitkov, in case we only computed anaphora related cases and disregarded those cases of "it" which were wrongly classified as expletives. Expletive "it" present in the text are 189: so at first we computed coverage and accuracy with the usual formula that we report below. Then we subtracted wrongly classified cases from the number of total "it" found in one case (following Mitkov who claims that wrongly classified "it" found by the system should not count; in another case, this number is subtracted from the total number of "it" to be found in the text. Only for MARS we then computed different measures of Coverage and Accuracy. If we regard this approach worth pursuing, we come up with two Adjusted Accuracy measures which are related to the revised total numbers of anaphors by the two subtractions indicated above.

We computed manually all third person pronominal expressions and came up with a figure 982 which is

Table 1. Expletive "it" compared results

	MARS	JavaRAP	GuiTAR	GETARUNS
Coverage	163 (86.2%)	188 (99.5%)	188 (99.5%)	171 (91%)
Accuracy 1	63 (33.3%)	73 (38.6%)	75 (39.7%)	87 (46 %)
Wrong Classification	44 163-44=119 189-44=145	49 189-49=140	64 189-64=125	53 189-53=136
Wrong Reference	56	66	49	32
Accuracy 2	63 (38.6%)			
Adjusted Accuracy 2	63 (52.9%)			
Adjusted Accuracy 3	63 (43.4%)	73 (52.1%)	75 (60%)	87 (64 %)

only confirmed by one of the three systems considered: JavaRAP. Pronouns considered are the following one, lower case and upper case included:

Possessives – his, its, her, hers, their, theirs

Personals – he, she, it, they, him, her, it, them (where "it" and "her" have to be disambiguated)

Reflexives – himself, itself, herself, themselves

There are 16 different wordforms. As can be seen from the table below, apart from JavaRAP, none of the other systems considered comes close to 100% coverage.

Computing general measures for Precision and Recall we have three quantities (see also Poesio & Kabadjov):

- total number of anaphors present in the text;
- anaphors identified by the system;
- correctly resolved anaphors.

Formulas related to Accuracy/Success Rate or Precision are as follows: Accuracy1 = number of successfully resolved anaphors/number of all anaphors; Accuracy2 = number of successfully resolved anaphors/number of anaphors found (attempted to be resolved). Recall - which should correspond to Coverage - we come up with formula: R= number of anaphors found /number of all anaphors to be resolved (present in the text). Finally the formula for F-measure is as follows: 2*P*R/(P+R) where P is chosen as Accuracy 2.

Table 2. Overall results Coverage/Accuracy

	COVERAGE	ACCURACY 1	ACCURACY 2	F-measure
MARS	936 (95.3%)	403/982 (41.5%)	403/903 (43%)	59.26%
JavaRAP	981 (100%)	490/982 (49.9%)	490/981 (50%)	66.7%
GUITAR	824 (84.8%)	445/982 (45.8%)	445/824 (54%)	65.98%
GETARUNS	885 (90.1%)	555/982 (56.5%)	555/885 (62.7%)	73.94%

In absolute terms best accuracy figures have been obtained by GETARUNS, followed by JavaRAP. So it is still thanks to the classic Recall formula that this result stands out clearly. We also produced another table which can however only be worked out for our system, which uses a distributed approach. We managed to separate pronominal expressions in relation to their contribution at the different levels of anaphora resolution considered: clause level, utterance level, discourse level. At clause level, only those pronouns which must be bound locally are checked, as is the case with reflexive pronouns, possessives, some cases of expletive 'it': both arguments and adjuncts may contribute the appropriate antecedent. At utterance level, in case the sentence is complex or there is more than one clause, also personal subject/object pronouns may be bound (if only preferentially so). Eventually, those pronouns which do not find an antecedent are regarded discourse level pronouns.

We collapsed under CLAUSE all pronouns bound at clause and utterance level; DISCOURSE contains only sentence external pronouns. Expletives have been computed in a separate column.

Table 3. GETARUNS pronouns collapsed at structural level

	CLAUSE	DISCOURSE	EXPLETIVES	TOTALS
Pronouns found	410	366	109	885
Correct	266	222	67	555
Errors made	144	144	42	330

As can be noticed easily, the highest percentage of pronouns found is at Clause level: this is not however the best performance of the system, which on the contrary performs better at discourse level. Expletives contribute by far the highest correct result. We also found correctly 47 'there' expletives and 6 correctly classified pronominal 'there' which however have been left unbound. The system also found 48 occurrences of deictic discourse bound "this" and "that", which corresponds to the full coverage.

Finally, nominal expressions: the History List (HL) has been incremented up to 2243 new entities. The system identified 2773 entities from the HL by matching their linguistic description. The overall number of resolution actions taken by the Discourse Level algorithm is 1861: this includes both cases of nominal and pronominal expressions. However, since only 366 can be pronouns, the remaining 1500 resolution actions have been carried out on nominal expressions present in the HL. If we compare these results to the ones computed by GuiTAR, which assign semantic indices to NamedEntities disregarding their status of anaphora, we can see that the whole text is made up of 12731 NEs. GuiTAR finds 1585 cases of identity relations between a NE and an antecedent. However, GuiTAR introduces always new indices and creates local antecedent-referring expression chains rather than repeating the same index of the chain head. In this way, it is difficult if not impossible to compute how many times the text corefers/cospecifies to the same referring expressions. On the contrary, in our case, this can be easily computed by counting how many times the same semantic index is being repeated in a "resolution" or "identity" action of the anaphora resolution algorithm. For instance, the Jury is coreferred/cospecified 12 times; Price Daniel also 12 times and so on.

5. Conclusions

The error rate of both Charniak's and Connexor's as reported in the literature, is approximately the same, 20%; this notwithstanding, MARS has a slightly reduced coverage when compared with JavaRAP, 96%. GuiTAR has the worst coverage, 85%. As to accuracy, none of the three algorithms overruns 50%: JavaRAP has the best score 49.9%. However GETARUNS has 63% correct score, with 90% coverage.

There are at least three reasons why our system has a better performance: one is the presence of a richer functional and semantic information as explained above, which comes with augmented head-dependent structures. Second reason is the decision to split the referential process into two and treat utterance level pronominal expressions separately from discourse level ones. Third reason is the way in which discourse level anaphora resolution is

organized: our version of the Centering algorithm hinges on a record of a list of best antecedents weighted on the basis of their behaviour in History List and on their intrinsic semantic properties. These three properties of our AR algorithm can be dubbed the Knowledge Rich approach.
F-measures approximates very closely what we obtained in a previous experiment: however, as a whole it is an insufficient score to insure adequate confidence in semantic substitution of anaphoric items by the head of the antecedent. Improvements need to come from parsing and the lexical component.

Acknowledgements

Thanks to three anonymous reviewers who helped us improve the overall layout of the paper.

References

Delmonte R. 1990. Semantic Parsing with an LFG-based Lexicon and Conceptual Representations, *Computers & the Humanities,* 5-6, pp.461-488.

Delmonte R. and D.Bianchi 1991. Binding Pronominals with an LFG Parser, *Proceeding of the Second International Workshop on Parsing Technologies*, Cancun(Messico), ACL 1991, pp.59-72.

Delmonte R., D.Bianchi 1992. Quantifiers in Discourse, in *Proc. ALLC/ACH'92*, Oxford(UK), OUP, pp. 107-114.

Delmonte R. 1992. *Linguistic and Inferential Processing in Text Analysis by Computer*, UP, Padova.

Delmonte R. and D.Bianchi 1994. Computing Discourse Anaphora from Grammatical Representation, in D.Ross & D.Brink(eds.), *Research in Humanities Computing* 3, Clarendon Press, Oxford, 179-199.

Delmonte R. and D.Bianchi 1999. Determining Essential Properties of Linguistic Objects for Unrestricted Text Anaphora Resolution, *Proc. Workshop on Procedures in Discours*e, Pisa, pp.10-24.

Delmonte R., L.Chiran, and C.Bacalu, (2000). Towards An Annotated Database For Anaphora Resolution, LREC, Atene, pp.63-67.

Delmonte R. 2002a. From Deep to Shallow Anaphora Resolution: What Do We Lose, What Do We Gain, in Proc. International Symposium RRNLP, Alicante, pp.25-34.

Delmonte R. 2002b. From Deep to Shallow Anaphora Resolution:, in *Proc. DAARC2002 , 4th Discourse Anaphora and Anaphora Resolution Colloquium*, Lisbona, pp.57-62.

Delmonte, R. 2003. Getaruns: a Hybrid System for Summarization and Question Answering. In Proc. Natural Language Processing (NLP) for Question-Answering, EACL, Budapest, pp. 21-28.

Delmonte R. 2004. Evaluating GETARUNS Parser with GREVAL Test Suite, In Proc. ROMAND - 20th COLING, University of Geneva, pp. 32-41.

Di Eugenio B. 1990. Centering Theory and the Italian pronominal system, COLING, Helsinki.

Grosz B. and C. Sidner 1986. Attention, Intentions, and the Structure of Discourse, *Computational Linguistics* 12 (3), 175-204.

Kennedy, C. and B. Boguraev, 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proc. of the 16th COLING*, Budapest.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua, 2004. A Public Reference Implementation of the RAP Anaphora Resolution Algorithm, In *Proceedings of the Language Resources and Evaluation Conference 2004 (LREC 04),* Lisbon, Portugal, pp.1-4.

Mitkov R. 1995. Two Engines are better than one: Generating more power and confidence in the search for the antecedent, *Proceedings of Recent Advances in Natural Language Processing*, Tzigov Chark, 87-94.

Mitkov, R. 1998. Robust Pronoun Resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 869-875, Montreal, Canada.

Mitkov, R., R. Evans, and C. Orasan. 2002. A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method, *Proceedings of CICLing*-2002, pp.1-19.

Poesio, M. and R. Vieira, 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Poesio, M. and Mijail A. Kabadjov 2004. A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation *Proceedings of the Language Resources and Evaluation Conference 2004 (LREC 04),* Lisbon, Portugal, pp.1-4.

Sidner C. 1983. Focusing in the Comprehension of Definite Anaphora, in Brady M., Berwick R.(eds.), *Computational Models of Discourse*, MIT Press, Cambridge, MA, 267-330.

Webber B. 1983. So can we Talk about Now?, in Brady M., Berwick R.(eds.), *Computational Models of Discourse*, MIT Press, Cambridge, MA, 331-371.

Webber B. L. 1991. Structure and Ostension in the Interpretation of Discourse Deixis, in *Language and Cognitive Processes* 6 (2):107-135.

Zribi-Hertz A. 1989. Anaphor Binding and Narrative Point of View: English reflexive pronouns in sentence and discourse, *Language*, 65(4):695-727.

Robust Parsing of the Proposition Bank

Gabriele Musillo
Depts of Linguistics and Computer Science
University of Geneva
2 Rue de Candolle
1211 Geneva 4
Switzerland
musillo4@etu.unige.ch

Paola Merlo
Department of Linguistics
University of Geneva
2 Rue de Candolle
1211 Geneva 4
Switzerland
merlo@lettres.unige.ch

Abstract

In this paper, we extend an existing statistical parsing model to produce richer output parse trees, annotated with PropBank semantic role labels. Our results show that the model can be robustly extended to produce more complex output parse trees without any loss in performance and suggest that joint inference of syntactic and semantic representations is a viable alternative to approaches based on a pipeline of local processing steps.

1 Introduction

Recent successes in statistical syntactic parsing based on supervised learning techniques trained on a large corpus of syntactic trees (Collins, 1999; Charniak, 2000; Henderson, 2003) have brought forth the hope that the same approaches could be applied to the more ambitious goal of recovering the propositional content and the frame semantics of a sentence. Moving towards a shallow semantic level of representation is a first initial step towards the distant goal of natural language understanding and has immediate applications in question-answering and information extraction. For example, an automatic flight reservation system processing the sentence *I want to book a flight from Geneva to Trento* will need to know that *from Geneva* denotes the origin of the flight and *to Trento* denotes its destination. Knowing that these two phrases are prepositional phrases, the information provided by a syntactic parser, is only moderately useful.

The growing interest in learning deeper information is to a large extent supported and due to the recent development of semantically annotated databases such as FrameNet (Baker et al., 1998) or the Proposition Bank (Palmer et al., 2005), that can be used as training resources for a number of supervised learning paradigms. We focus here on the Proposition Bank (PropBank). PropBank encodes propositional information by adding a layer of argument structure annotation to the syntactic structures of the Penn Treebank (Marcus et al., 1993). Verbal predicates in the Penn Treebank (PTB) receive a label REL and their arguments are annotated with abstract semantic role labels A0-A5 or AA for those complements of the predicative verb that are considered arguments while those complements of the verb labelled with a semantic functional label in the original PTB receive the composite semantic role label AM-X, where X stands for labels such as LOC, TMP or ADV, for locative, temporal and adverbial modifiers respectively. A tree structure with PropBank labels for a sentence from the PTB (section 00) is shown in Figure 1 below. PropBank uses two levels of granularity in its annotation, at least conceptually. Arguments receiving labels A0-A5 or AA do not express consistent semantic roles and are specific to a verb, while arguments receiving an AM-X label are supposed to be adjuncts and the respective roles they express are consistent across all verbs.[1]

Recent approaches to learning semantic role labels are based on two-stage architectures. The first stage selects the elements to be labelled, while the second determines the labels to be assigned to the selected elements. While some of these models are based on full parse trees (Gildea and Jurafsky, 2002; Gildea and Palmer, 2002), other methods have been proposed that eschew the need for a full

[1] There are thirteen semantic role labels for modifiers. See (Palmer et al., 2005) for a detailed discussion of PropBank semantic roles labels.

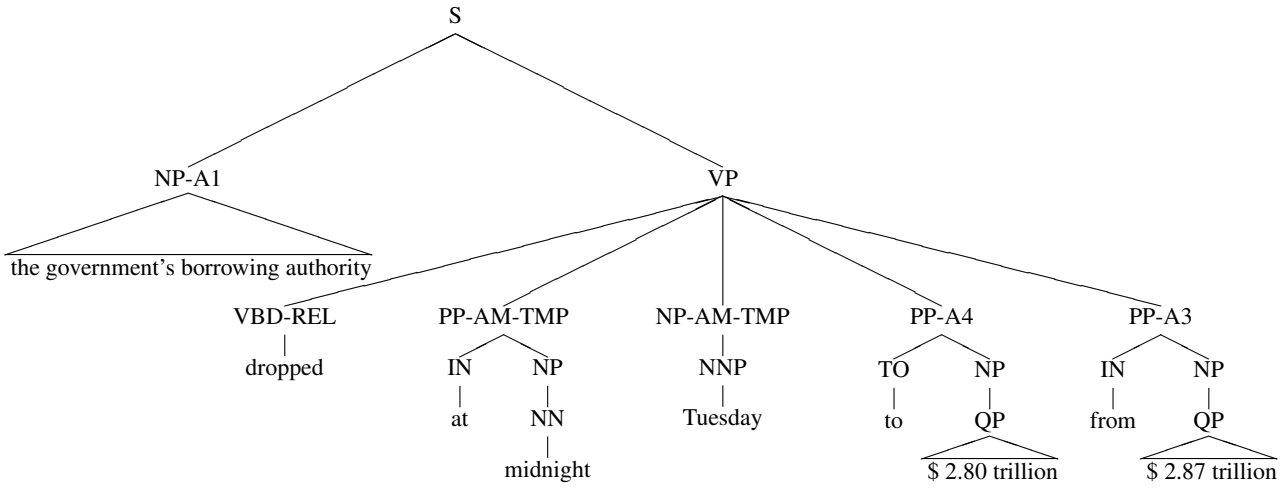


Figure 1: A sample syntactic structure from the PropBank with semantic role annotations.

parse (CoNNL, 2004; CoNLL, 2005). Because of the way the problem has been formulated – as a pipeline of parsing (or chunking) feeding into labelling – specific investigations of integrated approaches that solve both the parsing and the semantic role labelling problems at the same time have not been studied.

We present work to test the hypothesis that a current statistical parser (Henderson, 2003) can output richer information robustly, that is without any significant degradation of the parser's accuracy on the original parsing task, by explicitly modelling semantic role labels as the interface between syntax and semantics.

We achieve promising results both on the simple parsing task, where the accuracy of the parser is measured on the standard Parseval measures, and also on the parsing task where the more complex labels of PropBank are taken into account. We will call the former task Penn Treebank parsing (PTB parsing) and the latter task PropBank parsing below.

These results have several consequences. On the one hand, we show that it is possible to build a single integrated robust system successfully. This is a meaningful achievement, as a task combining semantic role labelling and parsing is more complex than simple syntactic parsing. While the shallow semantics of a constituent and its structural position are often correlated, they sometimes diverge. For example, some nominal temporal modifiers occupy an object position without being objects, like *Tuesday* in Figure 1 below. On the other hand, our results indicate that the proposed models are robust. To model our task accurately, additional parameters must be estimated. However, given the current limited availability of annotated treebanks, this more complex task will have to be solved with the same overall amount of data, aggravating the difficulty of estimating the model's parameters due to sparse data. The limited availability of data is increased further by the high variability of the argumental labels A0-A5 whose semantics is specific to a given verb or a given verb sense. Solving this more complex problem successfully, then, indicates that the models used are robust.

Finally, we achieve robustness without simplifying the parsing architecture. Specifically, robustness is achieved without resorting to the stipulation of strong independence assumptions to compensate for the limited availability and high variability of data. Consequently, such an achievement demonstrates not only that the robustness of the parsing model, but also its scalability and portability.

2 The Basic Parsing Architecture

To achieve the complex task of assigning semantic role labels while parsing, we use a family of statistical parsers, the Simple Synchrony Network (SSN) parsers (Henderson, 2003), which do not make any explicit independence assumptions, and are therefore likely to adapt without much modification to the current problem. This architecture has shown state-of-the-art performance.

SSN parsers comprise two components, one which estimates the parameters of a stochastic model for syntactic trees, and one which searches for the most probable syntactic tree given the

parameter estimates. As with many other statistical parsers (Collins, 1999; Charniak, 2000), SSN parsers use a history-based model of parsing. Events in such a model are derivation moves. The set of well-formed sequences of derivation moves in this parser is defined by a Predictive LR pushdown automaton (Nederhof, 1994), which implements a form of left-corner parsing strategy. The derivation moves include: projecting a constituent with a specified label, attaching one constituent to another, and shifting a tag-word pair onto the pushdown stack.

Unlike standard history-based models, SSN parsers do not state any explicit independence assumptions between derivation steps. They use a neural network architecture, called Simple Synchrony Network (Henderson and Lane, 1998), to induce a finite history representation of an unbounded sequence of moves. The history representation of a parse history $d_1, \ldots, d_{i-1}$, which we denote $h(d_1, \ldots, d_{i-1})$, is assigned to the constituent that is on the top of the stack before the ith move.

The representation $h(d_1, \ldots, d_{i-1})$ is computed from a set f of features of the derivation move d_{i-1} and from a finite set D of recent history representations $h(d_1, \ldots, d_j)$, where $j < i - 1$. Because the history representation computed for the move $i - 1$ is included in the inputs to the computation of the representation for the next move i, virtually any information about the derivation history could flow from history representation to history representation and be used to estimate the probability of a derivation move. However, the recency preference exhibited by recursively defined neural networks biases learning towards information which flows through fewer history representations. (Henderson, 2003) exploits this bias by directly inputting information which is considered relevant at a given step to the history representation of the constituent on the top of the stack before that step. In addition to history representations, the inputs to $h(d_1, \ldots, d_{i-1})$ include handcrafted features of the derivation history that are meant to be relevant to the move to be chosen at step i. For each of the experiments reported here, the set D that is input to the computation of the history representation of the derivation moves $d_1, \ldots, d_{i-1}$ includes the most recent history representation of the following nodes: top_i, the node on top of the pushdown stack before the ith move; the left-corner ancestor of top_i (that is, the second top-most node on the parser's stack); the leftmost child of top_i; and the most recent child of top_i, if any. The set of features f includes the last move in the derivation, the label or tag of top_i, the tag-word pair of the most recently shifted word, and the leftmost tag-word pair that top_i dominates. Given the hidden history representation $h(d_1, \cdots, d_{i-1})$ of a derivation, a normalized exponential output function is computed by SSNs to estimate a probability distribution over the possible next derivation moves d_i.[2]

The second component of SSN parsers, which searches for the best derivation given the parameter estimates, implements a severe pruning strategy. Such pruning handles the high computational cost of computing probability estimates with SSNs, and renders the search tractable. The space of possible derivations is pruned in two different ways. The first pruning occurs immediately after a tag-word pair has been pushed onto the stack: only a fixed beam of the 100 best derivations ending in that tag-word pair are expanded. For training, the width of such beam is set to five. A second reduction of the search space prunes the space of possible project or attach derivation moves: a best-first search strategy is applied to the five best alternative decisions only.

The next section describes our model, extended to produce richer output parse trees annotated with semantic role labels.

3 Learning Semantic Role Labels

Previous work on learning function labels during parsing (Merlo and Musillo, 2005; Musillo and Merlo, 2005) assumed that function labels represent the interface between lexical semantics and syntax. We extend this hypothesis to the semantic role labels assigned in PropBank, as they are an exhaustive extension of function labels, which have been reorganised in a coherent inventory of labels and assigned exhaustively to all sentences in the PTB. Because PropBank is built on the PTB, it inherits in part its notion of function labels which is directly integrated into the AM-X role labels. A0-A5 or AA labels correspond to many of the unlabelled elements in the PTB and also to those elements that PTB annotators had classified as re-

[2] The on-line version of Backpropagation is used to train SSN parsing models. It performs a gradient descent with a maximum likelihood objective function and weight decay regularization (Bishop, 1995).

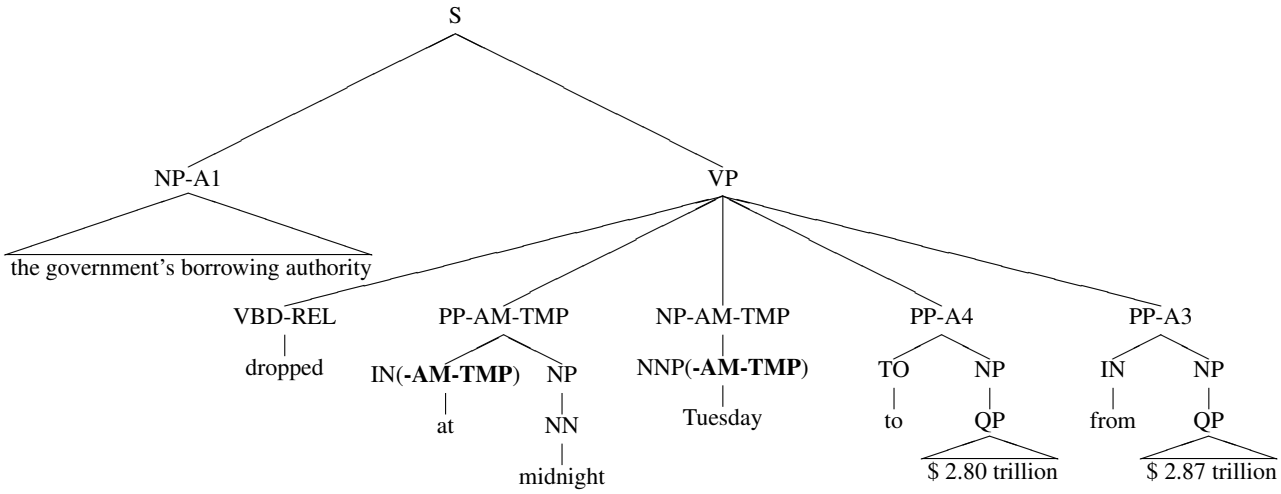


Figure 2: A sample syntactic structure with semantic role labels lowered onto the preterminals.

ceiving a syntactic functional label such as SBJ (subject) or DTV (dative).

Because they are projections of the lexical semantics of the elements in the sentence, semantic role labels are projected bottom-up, they tend to appear low in the tree and they are infrequently found on the higher levels of the parse tree, where projections of grammatical, as opposed to lexical, elements usually reside. Because they are the interface level with syntax, semantic labels are also subject to distributional constraints that govern syntactic dependencies, such as argument structure or subcategorization. We attempt to capture such constraints by modelling the c-command relation. Recall that the c-command relation relates two nodes in a tree, even if they are not close to each other, provided that the first node dominating one node also dominate the other. This notion of c-command captures both linear and hierarchical constraints and defines the domain in which semantic role labelling applies.

While PTB function labels appear to overlap to a large extent with PropBank semantic rolel labels, work by (Ye and Baldwin, 2005) on semantic labelling prepositional phrases, however, indicates that the function labels in the Penn Treebank are assigned more sporadically and heterogeneously than in PropBank. Apparently only the "easy" cases have been tagged functionally, because assigning these function tags was not the main goal of the annotation. PropBank instead was annotated exhaustively, taking all cases into account, annotating multiple roles, coreferences and discontinuous constituents. It is therefore not void of interest to test our hypothesis that, like function labels, semantic role labels are the interface between syntax and semantics, and they need to be recovered by applying constraints that model both higher level nodes and lower level ones.

We assume that semantic roles are very often projected by the lexical semantics of the words in the sentence. We introduce this bottom-up lexical information by fine-grained modelling of semantic role labels. Extending a technique presented in (Klein and Manning, 2003) and adopted in (Merlo and Musillo, 2005; Musillo and Merlo, 2005) for function labels, we split some part-of-speech tags into tags marked with semantic role labels. The semantic role labels attached to a non-terminal directly projected by a preterminal and belonging to a few selected categories (DIR, EXT, LOC, MNR, PNC, CAUS and TMP) were propagated down to the pre-terminal part-of-speech tag of its head. To affect only labels that are projections of lexical semantics properties, the propagation takes into account the distance of the projection from the lexical head to the label, and distances greater than two are not included. Figure 2 illustrates the result of this operation.

In our augmented model, inputs to each history representation are selected according to a linguistically motivated notion of structural locality over which dependencies such as argument structure or subcategorization could be specified.
In SSN parsing models, the set D of nodes that are structurally local to a given node on top of the stack defines the structural distance between this given node and other nodes in the tree. Such a notion of distance determines the number of history representations through which information passes

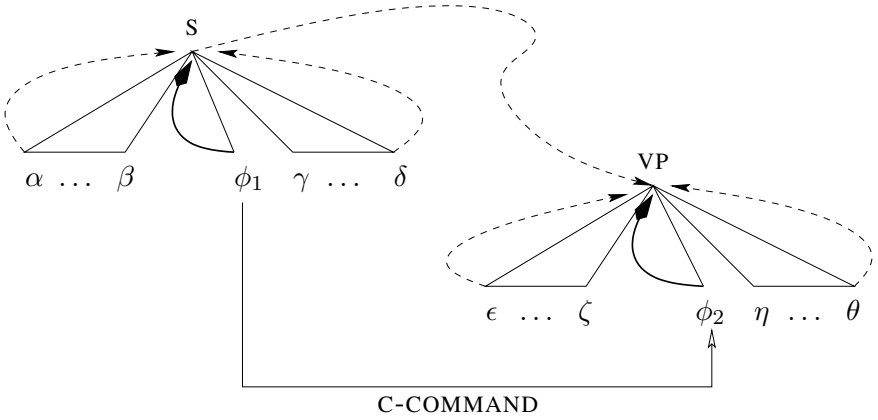


Figure 3: Flow of information in original SSN parsers (dashed lines), enhanced by biases specific to semantic role labels to capture the notion of c-command (solid lines).

to flow from the representation of a node i to the representation of a node j. By adding nodes to the set D, one can shorten the structural distance between two nodes and enlarge the locality domain over which dependencies can be specified. To capture a locality domain appropriate for semantic role parsing, we add the most recent child of top_i labelled with a semantic role label to the set D. These additions yield a model that is sensitive to regularities in structurally defined sequences of nodes bearing semantic role labels, within and across constituents. This modification of the biases is illustrated in Figure 3.

This figure displays two constituents, S and VP with some of their respective child nodes. The VP node is assumed to be on the top of the parser's stack, and the S one is supposed to be its left-corner ancestor. The directed arcs represent the information that flows from one node to another. According to the original SSN model in (Henderson, 2003), only the information carried over by the leftmost child and the most recent child of a constituent directly flows to that constituent. In the figure above, only the information conveyed by the nodes α and δ is directly input to the node S. Similarly, the only bottom-up information directly input to the VP node is conveyed by the child nodes ϵ and θ. In the original SSN models, nodes bearing a function label such as ϕ_1 and ϕ_2 are not directly input to their respective parents. In our extended model, information conveyed by ϕ_1 and ϕ_2 directly flows to their respective parents. So the distance between the nodes ϕ_1 and ϕ_2, which stand in a c-command relation, is shortened. For more information on this technique to capture domains induced by the c-command relation, see (Musillo and Merlo, 2005).

We report the effects of these augmentations on parsing results in the experiments described below.

4 Experiments

Our extended semantic role SSN parser was trained on sections 2-21 and validated on section 24 from the PropBank. Training, validating and testing data sets consist of the PTB data annotated with PropBank semantic roles labels, as provided in the CoNLL-2005 shared task (Carreras and Marquez, 2005).

Our augmented model has a total 613 of non-terminals to represents both the PTB and PropBank labels of constituents, instead of the 33 of the original SSN parser. The 580 newly introduced labels consist of a standard PTB label followed by a set of one or more PropBank semantic role such as PP-AM-TMP or NP-A0-A1. As a result of lowering the six AM-X semantic role labels, 240 new part-of-speech tags were introduced to partition the original tag set which consisted of 45 tags. SSN parsers do not tag their input sentences. To provide the augmented model with tagged input sentences, we trained an SVM tagger whose features and parameters are described in detail in (Gimenez and Marquez, 2004). Trained on section 2-21, the tagger reaches a performance of 95.45% on the test set (section 23) using our new tag set.
As already mentioned, argumental labels A0-A5 are specific to a given verb or a given verb sense, thus their distribution is highly variable. To reduce variability, we add some of the tag-verb pairs licensing these argumental labels to the vocabu-

	F	R	P
PropBank training and PropBank parsing task	82.3	82.1	82.4
PropBank training and PTB parsing task	88.8	88.6	88.9
PTB training and PTB parsing task (Henderson, 2003)	88.6	88.3	88.9

Table 1: Percentage F-measure (F), recall (R), and precision (P) of our SSN parser on two different tasks and the original SSN parser.

lary of our model. We reach a total of 4970 tag-word pairs.[3] This vocabulary comprises the original 512 pairs of the original SSN model, and our added pairs which must occur at least 10 times in the training data. Our vocabulary as well as the new 240 POS tags and the new 580 non-terminal labels are included in the set f of features input to the history representations as described in section 2.

We perform two different evaluations on our model trained on PropBank data. Recall that we distinguish between two parsing tasks: the PropBank parsing task and the PTB parsing task. To evaluate the first parsing task, we compute the standard Parseval measures of labelled recall and precision of constituents, taking into account not only the 33 original labels but also the 580 newly introduced PropBank labels. This evaluation gives us an indication of how accurately and exhaustively we can recover this richer set of non-terminal labels. The results, computed on the testing data set from the PropBank, are shown on the first line of Table 1.

To evaluate the PTB task, we compute the labelled recall and precision of constituents, ignoring the set of PropBank semantic role labels that our model assigns to constituents. This evaluation indicates how well we perform on the standard PTB parsing task alone, and its results on the testing data set from the PTB are shown on the second line of Table 1.

The third line of Table 1 gives the performance on the simpler PTB parsing task of the original SSN parser (Henderson, 2003), that was trained on the PTB data sets contrary to our SSN model trained on the PropBank data sets.

5 Discussion

These results clearly indicate that our model can perform the PTB parsing task at levels of performance comparable to state-of-the-art statistical parsing, by extensions that take the nature of the richer labels to be recovered into account. They also suggest that the relationship between syntactic PTB parsing and semantic PropBank parsing is strict enough that an integrated approach to the problem of semantic role labelling is beneficial.

In particular, recent models of semantic role labelling separate input indicators of the correlation between the structural position in the tree and the semantic label, such as *path*, from those indicators that encode constraints on the sequence, such as the previously assigned role (Kwon et al., 2004). In this way, they can never encode directly the constraining power of a certain role in a given structural position onto a following node in its structural position. In our augmented model, we attempt to capture these constraints by directly modelling syntactic domains defined by the notion of c-command.

Our results also confirm the findings in (Palmer et al., 2005). They take a critical look at some commonly used features in the semantic role labelling task, such as the *path* feature. They suggest that the path feature is not very effective because it is sparse. Its sparseness is due to the occurrence of intermediate nodes that are not relevant for the syntactic relations between an argument and its predicate. Our model of domains is less noisy, and consequently more robust, because it can focus only on c-commanding nodes bearing semantic role labels, thus abstracting away from those nodes that smear the pertinent relations.

(Yi and Palmer, 2005) share the motivation of our work. Like the current work, they observe that the distributions of semantic labels could potentially interact with the distributions of syntactic labels and redefine the boundaries of constituents, thus yielding trees that reflect generalisations over both these sources of information.

To our knowledge, no results have yet been published on parsing the PropBank. Accordingly, it is not possible to draw a straigthforward quantitative

[3] Such pairs consists of a tag and a word token. No attempt at collecting word types was made.

	F	R	P
(Haghighi et al., 2005)	83.4	83.1	83.7
(Pradhan et al., 2005)	83.3	83.0	83.5
(Punyakanok et al., 2005)	83.1	82.8	83.3
(Marquez et al., 2005)	83.1	82.8	83.3
(Surdeanu and Turmo, 2005)	82.7	82.5	83.0
PropBank SSN	81.6	81.3	81.9

Table 2: Percentage F-measure (F), recall (R), and precision (P) of our Propbank SSN parser and state-of-the-art semantic role labelling systems on the PropBank parsing task (1267 sentences from PropBank validating data sets; Propbank data sets are available at http://www.lsi.upc.edu/ srlconll/st05/st05.html).

comparison between our PropBank SSN parser and other PropBank parsers. However, state-of-the-art semantic role labelling systems (CoNLL, 2005) use parse trees output by state-of-the-art parsers (Collins, 1999; Charniak, 2000), both for training and testing, and return partial trees annotated with semantic role labels. An indirect way of comparing our parser with semantic role labellers suggests itself. We merge the partial trees output by a semantic role labeller with the output of a parser it was trained on, and compute PropBank parsing performance measures on the resulting parse trees. The first five lines of Table 2 report such measures for the five best semantic role labelling systems (Haghighi et al., 2005; Pradhan et al., 2005; Punyakanok et al., 2005; Marquez et al., 2005; Surdeanu and Turmo, 2005) according to (CoNLL, 2005). The partial trees output by these systems were merged with the parse trees returned by (Charniak, 2000)'s parser. These systems use (Charniak, 2000)'s parse trees both for training and testing as well as various other information sources including sets of n-best parse trees (Punyakanok et al., 2005; Haghighi et al., 2005) or chunks (Marquez et al., 2005; Pradhan et al., 2005) and named entities (Surdeanu and Turmo, 2005). While our preliminary results indicated in the last line of Table 2 are not state-of-the-art, they do demonstrate the viability of SSN parsers for joint inference of syntactic and semantic representations.

6 Conclusions

In this paper, we have explored extensions to an existing state-of-the-art parsing model. We have achieved promising results on parsing the Proposition Bank, showing that our extensions are sufficiently robust to produce parse trees annotated with shallow semantic information. Future work will lie in extracting semantic role relations from such richly annotated trees, for applications such as information extraction or question answering. In addition, further research will explore the relevance of semantic role features to parse reranking.

Acknowledgements

We thank the Swiss National Science Foundation for supporting this research under grant number 101411-105286/1. We also thank James Henderson and Ivan Titov for allowing us to use and modify their SSN software, Xavier Carreras for providing the CoNLL-2005 shared task data sets and the anonymous reviewers for their valuable comments.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics (ACL-COLING'98)*, pages 86–90, Montreal, Canada.

Christopher M. Bishop. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Xavier Carreras and Lluis Marquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of North American Chapter of Association for Computational Linguistics (NAACL'00)*, pages 132–139, Seattle, Washington.

Michael John Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. the-

sis, Department of Computer Science, University of Pennsylvania.

CoNLL. 2005. *Ninth Conference on Computational Natural Language Learning* (CoNLL-2005). Ann Arbor, MI, USA.

CoNNL. 2004. *Eighth Conference on Computational Natural Language Learning* (CoNLL-2004). Boston, MA, USA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 239–246, Philadelphia, PA.

Jesus Gimenez and Lluis Marquez. 2004. Svmtool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Aria Haghighi, Kristina Toutanova, and Christopher Manning. 2005. A joint model for semantic role labeling. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

James Henderson and Peter Lane. 1998. A connectionist architecture for learning to parse. In *Proceedings of 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pages 531–537, University of Montreal, Canada.

Jamie Henderson. 2003. Inducing history representations for broad-coverage statistical parsing. In *Proceedings of the Joint Meeting of the North American Chapter of the Association for Computational Linguistics and the Human Language Technology Conference (NAACL-HLT'03)*, pages 103–110, Edmonton, Canada.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the ACL (ACL'03)*, pages 423–430, Sapporo, Japan.

Namhee Kwon, Michael Fleischman, and Eduard Hovy. 2004. Senseval automatic labeling of semantic roles using maximum entropy models. In *Senseval-3*, pages 129–132, Barcelona, Spain.

Mitch Marcus, Beatrice Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

Lluis Marquez, Pere Comas, Jesus Gimenez, and Neus Catala. 2005. Semantic role labeling as sequential tagging. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

Paola Merlo and Gabriele Musillo. 2005. Accurate function parsing. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 620–627, Vancouver, British Columbia, Canada, October.

Gabriele Musillo and Paola Merlo. 2005. Lexical and structural biases for function parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 83–92, Vancouver, British Columbia, October.

Mark Jan Nederhof. 1994. *Linguistic Parsing and Program Transformations*. Ph.D. thesis, Department of Computer Science, University of Nijmegen.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105.

Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Semantic role chunking combining complementary syntactic views. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

Vasin Punyakanok, Peter Koomen, Dan Roth, and Wen tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

Mihai Surdeanu and Jordi Turmo. 2005. Semantic role labeling using complete syntactic analysis. In *Proceedings of CoNLL-2005*, Ann Arbor, MI USA.

Patrick Ye and Timothy Baldwin. 2005. Semantic role labelling of prepositional phrases. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages pp. 779–791, Jeju, South Korea.

Szu-ting Yi and Martha Palmer. 2005. The integration of semantic parsing and semantic role labelling. In *Proceedings of CoNLL'05*, Ann Arbor, Michigan.

A Robust and Efficient Parser for Non-Canonical Inputs

Philippe Blache
CNRS & Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence, France
pb@lpl.univ-aix.fr

Abstract

We present in this paper a parser relying on a constraint-based formalism called Property Grammar. We show how constraints constitute an efficient solution in parsing non canonical material such as spoken language transcription or e-mails. This technique, provided that it is implemented with some control mechanisms, is very efficient. Some results are presented, from the French parsing evaluation campaign *EASy*.

1 Introduction

Parsing spoken languages and non canonical inputs remains a challenge for NLP systems. Many different solutions have been experimented, depending on the kind of material to be parsed or the kind of application: in some cases, superficial information such as bracketing is enough whereas in other situations, the system needs more details. The question of robustness, and more generally the parsing strategy, is addressed differently according to these parameters. Classically, three families of solutions are proposed:

- Reducing the complexity of the output
- Controlling the parsing strategy
- Training and adapting the system to the type of input

In the first case, the idea consists in building structures with little information, even underspecified (which means the possibility of building partial structures). We find in this family the different shallow parsing techniques (see for example [Hindle83], [Abney96]). Unsurprisingly, the use of statistical methods is very frequent and efficient in this kind of application (see [Tjong Kim Sang00] for some results of a comparison between different shallow parsers). Generally, such parsers (being them symbolic or not) are deterministic and build non recursive units. In some cases, they can also determine relations between units.

The second family contains many different techniques. The goal is to control a given parsing strategy by means of different mechanisms. Among them, we can underline three proposals:

- Implementing recovering mechanisms, triggering specific treatments in case of error (cf. [Boulier05])
- Controlling the parsing process by means of probabilistic information (cf. [Johnson98])
- Controlling deep parsers by means of shallow parsing techniques (cf. [Crysmann02], [UszKoreit02], [Marimon02])

The last kind of control mechanism consists in adapting the system to the material to be parsed. This can be done in different ways:

- Adding specific information in order to reduce the search space of the parsing process. This kind of information can appear under the form of ad hoc rules or information depending on the kind of data to be treated.
- Adapting the resources (lexicon, grammars) to the linguistic material

These different strategies offer several advantages and some of them can be used together. Their interest is that the related questions of robustness and efficiency are both taken into account. However, they do not constitute a generic

solution in the sense that something has to be modified either in the goal, in the formalism or in the process. In other words, they constitute an additional mechanism to be plugged into a given framework.

We propose in this paper a parsing technique relying on a constraint-based framework being both efficient and robust without need to modify the underlying formalism or the process. The notion of constraints is used in many different ways in NLP systems. They can be a very basic filtering process as proposed by *Constraint Grammars* (see [Karlsson90]) or can be part to an actual theory as with *HPSG* (see [Sag03]), the *Optimality Theory* (see [Prince03]) or *Constraint Dependency Grammars* (cf. [Maruyama90]). Our approach is very different: all information is represented by means of constraints; they do not stipulate requirements on the syntactic structure (as in the above cited approaches) but represent directly syntactic knowledge. In this approach, robustness is intrinsic to the formalism in the sense that what is built is not a structure of the input (for example under the form of a tree) but a description of its properties. The parsing mechanism can then be seen as a satisfaction process instead of a derivational one. Moreover, it becomes possible, whatever the form of the input, to give its characterization. The technique relies on constraint relaxation and is controlled by means of a simple left-corner strategy. One of its interests is that, on top of its efficiency, the same resources and the same parsing technique is used whatever the input.

After a presentation of the formalism and the parsing scheme, we describe an evaluation of the system for the treatment of spoken language. This evaluation has been done for French during the evaluation campaign Easy.

2 Property Grammars: a constraint-based formalism

We present in this section the formalism of Property Grammars (see [Bès99] for preliminary ideas, and [Blache00], [Blache05] for a presentation). The main characteristics of Property Grammars (noted hereafter PG), is that all information is represented by means of constraints. Moreover, grammaticality does not constitute the core question but become a side effect of a more general notion called characterization: an input is not associated to a syntactic structure, but described with its syntactic properties.

PG makes it possible to represent syntactic information in a decentralized way and at different levels. Instead of using sub-trees as with classical generative approaches, PG specifies directly constraints on features, categories or set of categories, independently of the structure to which they are supposed to belong. This characteristic is fundamental in dealing with partial, underspecified or non canonical data. It is then possible to stipulate relations between two objects, independently from their position in the input or into a structure. The description of the syntactic properties of an input can then be done very precisely, including the case of non canonical or non grammatical input. We give in the remaining of the section a brief overview of GP characteristics

All syntactic information is represented in PG by means of constraints (also called properties). They stipulate different kinds of relation between categories such as linear precedence, imperative co-occurrence, dependency, repetition, etc. There is a limited number of types of properties. In the technique described here, we use the following ones:

- Linear precedence: $Det < N$ (a determiner precedes the noun)
- Dependency: $AP \rightarrow N$ (an adjectival phrase depends on the noun)
- Requirement: $V[inf] \Rightarrow to$ (an infinitive comes with *to*)
- Exclusion: $seems \neq ThatClause[subj]$ (the verb *seems* cannot have *That* clause subjects)
- Uniqueness : $Uniq_{NP}\{Det\}$(the determiner is unique in a *NP*)
- Obligation : $Oblig_{NP}\{N, Pro\}$(a pronoun or a noun is mandatory in a *NP*)

This list can be completed according to the needs or the language to be parsed. In this formalism, a category, whatever its level is described with a set of properties, all of them being at the same level and none having to be verified before another.

Parsing a sentence in PG consists in verifying for each category the set of corresponding properties in the grammar. More precisely, the idea consists in verifying for each constituent subset its relevant constraints (i.e. the one applying to the ele-

ments of the subset). Some of these properties can be satisfied, some other can be violated. The result of this evaluation, for a category, is a set of properties together with their evaluation. We call such set the characterization of the category. Such an approach makes it possible to describe any kind of input.

Such flexibility has however a cost: parsing in PG is exponential (cf. [VanRullen05]). This complexity comes from several sources. First, this approach offers the possibility to consider all categories, independently from its corresponding position in the input, as possible constituent for another category. This makes it possible for example to take into account long distance or non projective dependencies between two units. Moreover, parsing non canonical utterances relies on the possibility of building characterizations with satisfied and violated constraints. In terms of implementation, a property being a constraint, this means the necessity to propose a constraint relaxation technique. Constraint relaxation and discontinuity are the main complexity factors of the PG parsing problem. The technique describe in the next section propose to control these aspects.

3 Parsing in PG

Before a description of the controlled parsing technique proposed here, we first present the general parsing schemata in PG. The process consists in building the list of all possible sets of categories that are potentially constituents of a syntactic unit (also called *constructions*). A characterization is built for each of this set. Insofar as constructions can be discontinuous, it is necessary to build all possible combinations of categories, in other words, the subsets set of the categories corresponding to the input to be parsed, starting from the lexical categories. We call *assignment* such a subset. All assignments have then, theoretically, to be evaluated with respect to the grammar. This means, for each assignment, traversing the constraint system and evaluating all relevant constraints (i.e. constraints involving categories belonging to the assignment). For some assignments, no property is relevant and the corresponding characterization is the empty set: we say in this case that the assignment in *non productive*. In other cases, the characterization is formed with all the evaluated properties, whatever their status (satisfied or not). At the first stage, all constructions contain only lexical categories, as in the following example:

Construction	*Assignment*	*Characterization*
AP	{Adv, Adj}	{Adv < Adj; Adv → Adj; ...}
NP	{Det, N}	{Det < N; Det → N; N ≠ Pro; ...}

An assignment with a productive characterization entails the instantiation of the construction as a new category; added to the set of categories. In the previous examples, *AP* and *NP* are then added to the initial set of lexical categories. A new set of assignments is then built, including these new categories as possible constituents, making it possible to identify new constructions. This general mechanism can be summarized as follows:

```
Initialization
    ∀ word at a position i:
       create the set c_i of its possible
       categories
    K ←  {c_i | 1<i<number of words}
    S ←  set of subsets of K
Repeat
    ∀ S_i ∈ S
       if S_i  is a productive assignment
         add k_i the characterization
              label to K
    S ←  set of subsets of K
Until new characterization are built
```

This parsing process underlines the complexity coming from the number of assignments to be taken into account: this set has to be rebuilt at each step (i.e. when a new construction is added).

As explained above, each assignment has to be evaluated. This process comes to build a characterization formed by the set of its relevant properties. A property p is relevant for an assignment A when A contains categories involved in the evaluation of p. In the case of unary properties constraining a category c, the relevance is directly known. In the case of n-ary properties, the situation is different for positive or negative properties. The former (e.g. cooccurrence constraints) concern two realized categories. In this case, c_1 and c_2 being these categories, we have $\{c_1, c_2\} \subset A$. In the case of negative properties (e.g. cooccurrence restriction), we need to have either $c_1 \notin A$ or $c_2 \notin A$.

When a property is relevant for a given A, its satisfiability is evaluated, according to the prop-

erty semantics, each property being associated to a solver. The general process is described as follows:

```
Let G the set of properties in the gram-
mar, let A an assignment

∀ p_i ∈ G, if p_i is relevant
     Evaluate the satisfiability of p_i
       for A
     Add p_i and its evaluation to the
       characterization C of A
Check whether C is productive
```

In this process, for all assignments, all properties have to be checked to verify their relevance and eventually their satisfiability.

The last aspect of this general process concerns the evaluation of the productivity of the characterization or an assignment. A productive assignment makes it possible to instantiate the corresponding category and to consider it as realized. A characterization is obviously productive when all properties are satisfied. But it is also possible to consider an assignment as productive when it contains violated properties. It is then possible to build categories, or more generally constructions, even for non canonical forms. In this case, the characterization is not entirely positive. This process has to be controlled. The basic control consists in deciding a threshold of violated constraints. It is also possible to be more precise and propose a hierarchization of the constraint system: some types of constraints or some constraints can play a more important role than others (cf. [Blache05b]).

A controlled version of this parsing schema, implemented in the experimentation described in the next section, takes advantage of the general framework, in particular in terms of robustness implemented as constraint relaxation. The process is however controlled for the construction of the assignment.

This control process relies on a left-corner strategy, adapted to the PG parsing schema. This strategy consists in identifying whether a category can start a new phrase. It makes it possible to drastically reduce the number of assignments and then control ambiguity. Moreover, the left corner suggests a construction label. The set of properties taken into consideration when building the characterization is then reduced to the set of properties corresponding to the label. These two controls, plus a disambiguation of the lexical level by means of an adapted POS tagger, render the parsing process very efficient.

The left corner process relies on a *precedence table*, calculated for each category according to the precedence properties in the grammar. This table is built automatically in verifying for each category whether, according to a given construction, it can precede all the other categories. The process consists in verifying that the category is not a left member of a precedence property of the construction. If so, the category is said to be a possible left corner of the construction. The precedence table contains then for each category the label of the construction for which it can be left corner.

During the process, when a category is a potential left corner of a construction *C*, we verify that the *C* is not the last construction opened by a left corner. If so, a new left corner is identified, and *C* is added to the set of possible constituents (usable by other assignments). Moreover, the characterization of the assignment beginning with c_i is built in verifying the subset of properties describing *C*.

The generation of the assignments can also be controlled by means of a co-constituency table. This table consists for each category, in indicating all the categories with which it belongs to a positive property. This table is easily built with a simple traversal of the constraint system. Adding a new category c_i to an assignment *A* is possible only when c_i appears as a co-constituent of a category belonging to *A*.

```
S initial set of lexical categories
Identification all the left corners
For all C, construction opened by a left
       corner c_i with G' the set of
       properties describing C
  Build assignments beginning by c_i
  Build characterizations verifying G'
```

The parsing mechanism described here takes advantage of the robustness of PG. All kind of input, whatever its form, can be parsed because if the possibility of relaxing constraints. Moreover, the control technique makes it possible to reduce the complexity of the process without modifying its philosophy.

4 Evaluation

We experimented this approach during the French evaluation campaign EASy (cf. [Paroubek05]). The test consisted in parsing several files containing various kinds of material: literature, newspaper, technical texts, questions, e-mails and spoken language. The total size of this corpus is one million words. Part of this corpus was annotated with morpho-syntactic (POS tags) and syntactic annotations. The last one provides bracketing as well as syntactic relations between units. The annotated part of the corpus represents 60,000 words and constitutes the gold standard.

The campaign consisted for the participants to parse the entire corpus (without knowing what part of the corpus constituted the reference). The results of the campaign are not yet available concerning the evaluation of the relations. The figures presented in this section concern constituent bracketing. The task consisted in identifying minimal non recursive constituents described by annotation guidelines given to the participants. The different categories to be built are: GA (adjective group: adjective or passed participle), GN (nominal group: determiner, noun adjective and its modifiers), GP (prepositional group), GR (adverb), NV (verbal nucleus: verb, clitics) and PV (verbal propositional group).

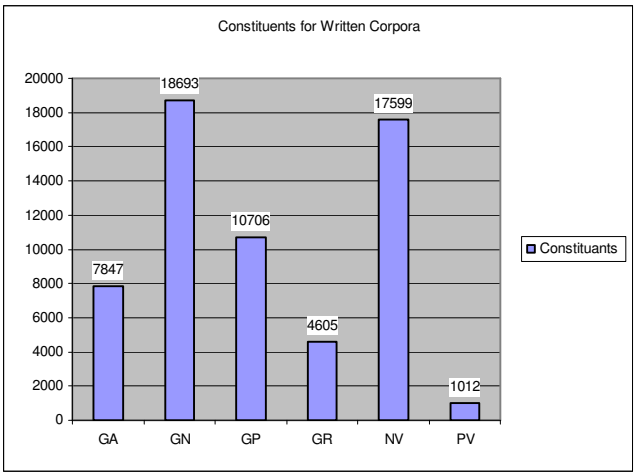
Our system parses the entire corpus (1 million words) in 4 minutes on a PC. It presents then a very good efficiency.

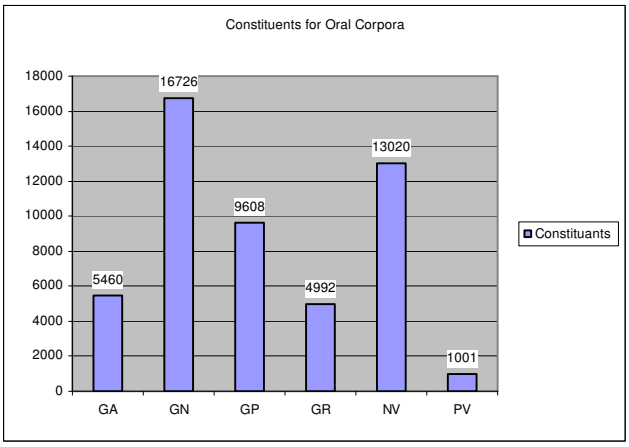
We have grouped the different corpora into three different categories: written texts (including newspapers, technical texts and literature), spoken language (orthographic transcription of spontaneous speech) and e-mails. The results are the following:

	Precision	*Recall*	*F-mesure*
Written texts	77.78	82.96	79.84
Spoken language	75.13	78.89	76.37
E-Mails	71.86	79.06	74.42

These figures show then very stable results in precision and recall, with only little loss of efficiency for non-canonical material. When studying more closely the results, some elements of explanation can be given. The e-mail corpus is to be analyzed separately: many POS tagging errors, due to the specificity of this kind of input explain the difference. Our POS-tagger was not tuned for this kind of lexical material.

The interpretation of the difference between written and oral corpora can have some linguistic basis. The following figures give quantitative indications on the categories built by the parser. The first remark is that the repartition between the different categories is the same. The only main difference concerns the higher number of nucleus VP in the case of written texts. This seems to support the classical idea that spoken language seems to use more nominal constructions than the written one.





The problem is that our parser encounters some difficulties in the identification of the NP borders. It very often also includes some material belonging in the grammar given during the campaign to AP or VP. The higher proportion of NPs in spoken corpora is an element of explanation for the difference in the results.

5 Conclusion

The first results obtained during the evaluation campaign described in this paper are very interesting. They illustrate the relevance of using symbolic approaches for parsing non-canonical material. The technique described here makes it possible to use the same method and the same resources whatever the kind of input and offers the possibility to do chunking as well as deep analysis. Moreover, such techniques, provided that they are implemented with some control mechanisms, can be very efficient: our parser treat more than 4,000 words per second. It constitutes then an efficient tool capable of dealing with large amount of data. On top of this efficiency, the parser has good results in terms of bracketing, whatever the kind of material parsed. This second characteristics also shows that the system can be used in real life applications.

In terms of theoretical results, such experimentation shows the interest of using constraints. First, they makes it possible to represent very fine-level information and offers a variety of control mechanisms, relying for example on the possibility of weighting them. Moreover, constraint relaxation techniques offer the possibility of building categories violating part of syntactic description of the grammar. They are then particularly well adapted to the treatment of non canonical texts. The formalism of Property Grammars being a fully constraint-based approach, it constitutes an efficient solution for the description of any kind of inputs.

Reference

[Abney 96] Abney S. (1996) "Partial Parsing via Finite-State Calculus", in proceedings of ESSLLI'96 Robust Parsing Workshop

[Bès99] Bès G. (1999) "La phrase verbale noyau en français", in *Recherches sur le français parlé*, 15, Université de Provence.

[Blache00] Blache P. (2000) "Constraints, Linguistic Theories and Natural Language Processing", in *Natural Language Processing*, D. Christodoulakis (ed), LNAI 1835, Springer-Verlag

[Blache05a] Blache P. (2005) "Property Grammars: A Fully Constraint-Based Theory", in *Constraint Solving and Language Processing,* H. Christiansen & al. (eds), LNAI 3438, Springer

[Boullier 05] Boullier P. & B. Sagot (2005) "Efficient and robust LFG parsing: SxLfg", in Proceedings of *IWPT '05*.

[Crysmann02] Crysmann B. A. Frank, B. Kiefer, S. Müller, G. Neumann, J. Piskorski, U. Schäfer, M. Siegel, H. Uszkoreit, F. Xu, M. Becker & H. Krieger (2002) "An Integrated Architecture for Shallow and Deep Processing", in proceedings of *ACL-02*.

[Frank03] Frank A., M. Becker, B. Crysmann, B. Kiefer & U. Schäfer (2003) "Integrated Shallow and Deep Parsing: TopP meets HPSG", in proceedings of *ACL-03*.

[Hindle83] Hindle D. (1983) *User manual for Fidditch, a deterministic parser*, Technical memorandum 7590-142, Naval Research Laboratory.

[Johnson98] Johnson M. (1998) "PCFG Models of Linguistic Tree Representations"', in *Computational Linguistics*, 24:4.

[Karlsson90] Karlsson F. (1990) "Constraint grammar as a framework for parsing running texts", in proceedings of *ACL-90*.

[Marimon02] Marimon M. (2002) "Integrating Shallow Linguistic Processing into a Unification-Based Spanish Grammar", in proceedings of *COLING-02*.

[Maruyama90] Maruyama H. (1990) "Structural Disambiguation with Constraint Propagation"', in proceedings of ACL'90.

[Paroubek05] Paroubek P., L. Pouillot, I. Robba & A. Vilnat (2005) "EASy : campagne d'évaluation des analyseurs syntaxiques", in proceedings of the workshop *EASy, TALN-2005*.

[Prince93] Prince A. & Smolensky P. (1993) "Optimality Theory: Constraint Interaction in Generative Grammars", Technical Report RUCCS TR-2, Rutgers Center for Cognitive Science.

[Tjong Kim Sang00] Tjong Kim Sang E. & S Buchholz (2000) "Introduction do the CoNLL-2000 Shared Task: Chunking", in proceedings of *CoNLL-2000*.

[Uszkoreit02] Uszkoreit H. (2002) "New Chances for Deep Linguistic Processing", in proceedings of *COLING-02*.

[VanRullen05] Van Rullen T. (2005), *Vers une analyse syntaxique à granularité variable*, PhD Thesis, Université de Provence.

Robust Parsing: More with Less

Kilian Foth, Wolfgang Menzel
Fachbereich Informatik, Universität Hamburg, Germany
foth|menzel@informatik.uni-hamburg.de

Abstract

Covering as many phenomena as possible is a traditional goal of parser development, but the broader a grammar is made, the blunter it may become, as rare constructions influence the behaviour on simple sentences that were already solved correctly. We observe the effects of intentionally removing support for specific constructions from a broad-coverage grammar of German. We show that accuracy of analysing sentences from the NEGRA corpus can be improved not only for sentences that do not need the extra coverage, but even when including those that do.

1 Introduction

Traditionally, broad coverage has always been considered to be a desirable property of a grammar: the more linguistic phenomena are treated properly by the grammar, the better results can be expected when applying it to unrestricted text (c.f. (Grover et al., 1993; Doran et al., 1994)). With the advent of empirical methods and the corresponding evaluation metrics, however, this view changed considerably. (Abney, 1996) was among the first who noted that the relationship between coverage and statistical parsing quality is a more complex one. Adding new rules to the grammar, i.e. increasing its coverage, does not only allow the parser to deal with more phenomena, hence more sentences; at the same time it opens up new possibilities for abusing the newly introduced rules to mis-analyse constructions which were already treated properly before. As a consequence, a net reduction in parsing quality might be observed for simple statistical reasons, since the gain usually is obtained for relatively rare phenomena, while the adverse effects might well affect frequent ones.

(Abney, 1996) uses this observation to argue in favour of stochastic models which attempt to choose the optimal structural interpretation instead of only providing a list of equally probable alternatives. However, using such an optimization procedure is not necessarily a sufficient precondition to completely rule out the effect. Compared to traditional handwritten grammars, successful stochastic models like (Collins, 1999; Charniak, 2000) open up an even greater space of alternatives for the parser and accordingly offer a great deal of opportunities to construct odd structural descriptions from them. Whether the guidance of the stochastic model can really prevent the parser from making use of these unwanted opportunities so far remains unclear.

In the following we make a first attempt to quantify the consequences that different degrees of coverage have for the output quality of a wide-coverage parser. For this purpose we use a Weighted Constraint Dependency Grammar (WCDG), which covers even relatively rare syntactic phenomena of German and performs reliably across a wide variety of different text genres (Foth et al., 2005). By combining hand-written rules with an optimization procedure for hypothesis selection, such a parser makes it possible to successively exclude certain rare phenomena from the coverage of the grammar and to study the impact of these modifications on its output quality

2 Some rare phenomena of German

What are good candidates of ‘rare’ phenomena that might be intentionally removed from the coverage of our grammar? One possibility is to remove coverage for constructions that are already slightly dispreferred. For instance, apposition and coordination of noun phrases often violate the principle of projectivity:

“I got a sled for Christmas, a parrot and a motor-bike.”

This is quite a common construction, but still ‘rare’ in the sense that the great majority of appositions does respect projectivity, so that the example seems at least slightly unusual. But there are also syntactic relations that are quite rare but nevertheless appear perfectly normal when they do occur, such as direct appellations:

“James, please open the door.”

This might be because their frequency varies considerably between text types; everyone is familiar with personal appellation from everyday conversation, but it would be surprising to hear it from the mouth of a television news reader.

Finally, some constructions form variants e.g. by omitting certain words:

“I bought a new broom [in order] to clean the drive-

No.	Phenomenon	Example	$f/1000$
1	*Mittelfeld* extraposition	"Es strahlt über DVB-T neben dem Fernsehprogramm auch **seinen Digitext** aus, **einen Videotext-ähnlichen Informationsdienst**."	32.5
2	ethical dative	"Noch erobere **sich** der PC neue Käuferschichten, heißt es weiter."	18.5
3	Nominalization	"Täglich kommen rund **1000 neue** hinzu."	13.4
4	Vocative	"So nicht, **ICANN**!"	9.7
5	Parenthetical matrix clause	"Bis zum Jahresende 2002, **prognostiziert Roland Berger**, werden die am Neuen Markt gelisteten Unternehmen 200.000 Mitarbeiter beschäftigen."	8.8
6	verb-first subclause	"**Erfüllt ein Mitgliedstaat keines oder nur eines dieser Kriterien**, so erstellt die Kommission einen Bericht."	8.3
7	Headline phrase	"**Lehrer** kaum auf Computer **vorbereitet**"	3.9
8	coordination cluster	"Auf den Webseiten der Initiative können **Spender PCs anbieten und Schulen ihren Bedarf anmelden**."	3.1
9	Adverbial pronoun	"Ihre Sprachen sollen **alle** gleichberechtigt sein."	2.6
10	*um* omission	"Und Dina ging aus, **die Töchter des Landes zu sehen**."	2.1
11	Metagrammatical usage	"Die Bezugnahmen auf die gemeinsame Agrarpolitik oder auf die Landwirtschaft und die Verwendung des Wortes **"landwirtschaftlich"** sind in dem Sinne zu verstehen, dass damit unter Berücksichtigung der besonderen Merkmale des Fischereisektors auch die Fischerei gemeint ist."	1.8
12	Auxiliary flip	"Die Geschädigten werfen Ricardo nun eine erhebliche Mitschuld vor, da größerer Schaden **hätte verhindert werden können**, wenn der Anbieter sofort gesperrt worden wäre."	1.1
13	Adjectival subclause	"Die Union unterhält ferner, **soweit zweckdienlich**, Beziehungen zu anderen internationalen Organisationen."	0.9
14	Suffix drop	"Ein **freundlich** Wort, das Maslo intervenieren ließ:"	0.5
15	Elliptical genitive	"**Martins** war auch nicht besser."	0.3
16	Adverbial noun	"Sie stehen sich **Auge in Auge** gegenüber."	0.1
17	Verb/particle mismatch	"Außer Windows 9x selbst **können** auch andere Hard- und Softwarekomponenten eines PC mit zu viel Hauptspeicher manchmal nicht **zurecht**."	0.1
18	*Vorfeld* extraposition	"**Der Verdacht** liegt nahe, **daß hier Schwarzarbeit betrieben wird**."	0.1
19	double relative subject	"Ich bin der Herr, **der ich** dich aus Ägyptenland herausgeführt habe."	0.02
20	Relative subject clause	"**Die dir fluchen**, seien verflucht, und **die dich segnen**, seien gesegnet!"	0.04
21	NP extraposition	"Die Verpflichtungen und die Zusammenarbeit in diesem Bereich bleiben im Einklang mit den im Rahmen **der Nordatlantikvertrags-Organisation** eingegangenen Verpflichtungen, **die für die ihr angehörenden Staaten weiterhin das Fundament ihrer kollektiven Verteidigung und das Instrument für deren Verwirklichung ist**."	0.01

Table 1: Some rare phenomena in modern German.

way."

Here the longer variant is unambiguously a subclause expressing purpose, while the shorter might be mistaken for a prepositional phrase, so it could be regarded as misleading for the parser.

The selection is necessarily subjective, not only because the delimitation of a phenomenon is subjective (are all kinds of ellipsis fundamentally the same phenomenon or not?) but also because we can remove only those phenomena that are already covered in the first place. Therefore we have selected phenomena

- that were explicitly added to the grammar at some point in order to deal with actually occurring unforeseen constructions,
- that can easily be removed from the grammar without affecting other phenomena,
- and that are relatively rare in all the texts we have investigated.

Table 1 shows the 21 phenomena that we consider in this paper. (Note that the three earlier example sentences correspond to lines 1, 4, and 10 in this table, but that not all lines have exact counterparts in English.) The last column gives the overall frequency per 1,000 sentences of each phenomenon when measured across all trees in our collection.

The collection contains sections of Bible text (Genesis 1–50), law text (the constitutions of Federal Germany and of the European Union), online technical newscasts (`www.heise.de`), novel text, and sentences from the NEGRA corpus of newspaper articles. Table 2 shows the sentence counts of the different sections and the frequency per 1000 of all 21 phenomena in each text type. It can be seen that most of the constructions remain quite rare overall, but often the frequency depends heavily on the text type, so that a high influence of the corpus can be expected for our experiments.

f/1000 Phen.	Bible (2,709)	Law (3,722)	Online (55,327)	Novel (20,253)	News (4,000)	overall (86,011)
1	93.6	24.6	29.0	36.7	28.2	32.6
2	59.6	17.5	12.2	31.3	16.2	18.6
3	21.0	22.7	12.3	12.4	19.5	13.4
4	18.4	0.0	0.1	38.2	1.2	9.7
5	1.1	0.0	5.8	18.2	15.8	8.8
6	3.4	51.4	7.8	2.6	6.8	8.3
7	0.7	3.6	4.8	1.3	7.2	3.9
8	7.1	4.4	3.3	2.4	1.8	3.1
9	7.1	0.5	1.6	5.0	3.5	2.6
10	12.7	1.9	1.9	1.2	1.2	2.0
11	0.4	0.3	2.2	0.5	4.8	1.8
12	1.5	0.0	0.9	1.8	1.5	1.1
13	2.2	0.8	1.0	0.5	0.2	0.9
14	0.7	0.0	0.6	1.2	0.2	0.7
15	1.9	0.0	0.7	0.0	1.0	0.5
16	0.4	0.3	0.2	0.0	0.0	0.1
17	1.1	0.0	0.1	0.0	0.0	0.1
18	0.0	0.0	0.1	0.0	0.2	0.1
19	0.7	0.0	0.0	0.0	0.2	0.0
20	0.7	0.0	0.0	0.0	0.0	0.0
21	0.0	0.3	0.0	0.0	0.0	0.0

Table 2: Frequency of phenomena by text type.

3 Weighted Constraint Dependency Grammar

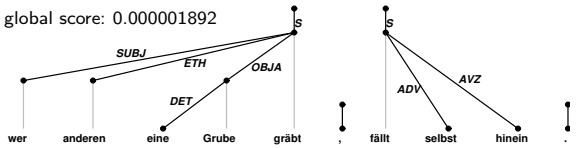
In WCDG (Schröder, 2002), natural language is modelled as labelled *dependency trees*, in which each word is assigned exactly one other word as its regent (only the root of the syntax tree remains unsubordinated) and a label that describes the nature of their relation. The set of acceptable trees is defined not by way of *generative* rules, but only through *constraints* on well-formed structures. Every possible dependency tree is considered correct unless one of its edges or edge pairs violates a constraint. This permissiveness extends to many properties that other grammar formalisms consider non-negotiable; for instance, a WCDG can allow non-projective (or, indeed, cyclical) dependencies simply by not forbidding them. Since the constraints can be arbitrary logical formulas, a grammar rule can also allow some types of non-projective relations and forbid others, and in fact the grammar in question does precisely that.

Weighted constraints can be written to express the fact that a construction is considered acceptable but not fully so. This mechanism is used extensively to achieve robustness against proper errors such as wrong inflection, ellipsis or mis-ordering; all of these are in fact expressed through defeasible constraints. But it can also express more subtle dispreferences against a specific phenomenon by writing only a weak constraint that forbids it; most of the phenomena listed in Table 1 are associated with such constraints to ensure that the parser assumes a rare construction only when this is necessary.

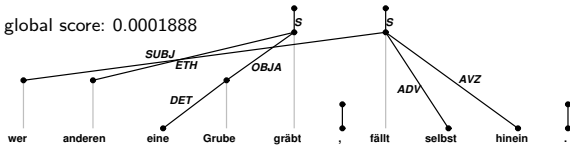
We employ a previously existing wide-coverage WCDG of modern German (Foth et al., 2005) that covers all of the presented rare phenomena. It comprises about 1,000 constraints, 370 of which are hard constraints. The entire parser and the grammar of German are publicly available at `http://nats-www.informatik.uni-hamburg.de/Papa/PapaDownloads`.

The optimal structure could be defined as the tree that violates the least important constraint (as in Optimality Theory), or the tree that violates the fewest constraints; in fact a multiplicative measure is used that combines both aspects by minimizing the collective dispreference for all phenomena in a sentence. Unfortunately, the resulting combinatorial problem is $\mathcal{NP}$-complete and admits of no efficient exact solution algorithm. However, variants of a heuristic *local search* can be used, which try to find the optimal tree by constructing a complete tree and then changing it in those places that violate important constraints. This involves a trade-off between parsing accuracy and processing time, because the correct structure is more likely to be found if there is more time to try out more alternatives. Given enough time, the method works well enough that the overall system exhibits a competitive accuracy even though the theoretical accuracy of the language model may be compromised by search errors.

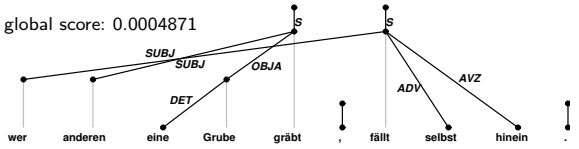
As an example of the process, consider the following analysis of the German proverb "Wer anderen eine Grube gräbt, fällt selbst hinein." *(He who digs a hole for others, will fall into it himself.)* The transformation starts with the following initial assumption



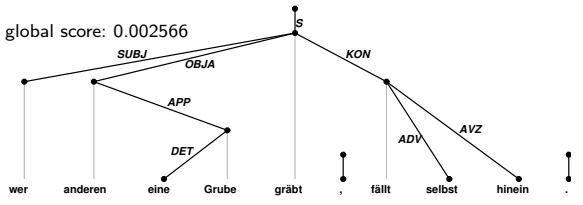
which, besides producing two isolated fragments instead of a spanning tree, also lacks a subject for the second clause.



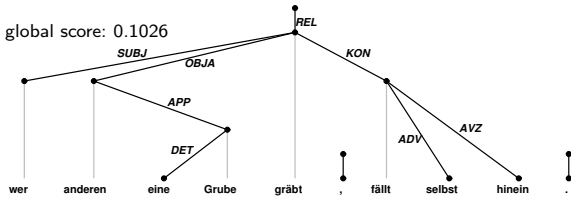
To mend this problem the relative pronoun from the first clause has been taken as a subject for the second one, with the result that the conflict has simply been moved to the first part of the sentence. Nevertheless, the global score improved considerably, since the verb-second condition for German main clauses is violated less often.



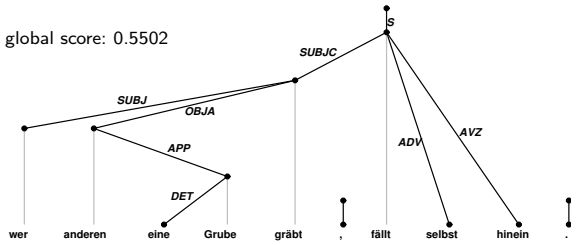
Here, the indefinite plural pronoun 'anderen' is taken as the subject for the second clause, creating, however, an agreement error with the finite verb, which is singular. Both subclauses have still not been integrated into a single spanning tree.



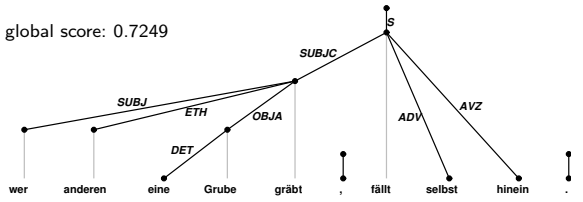
The integration is then achieved, but unfortunately as a coordination without an appropriate conjunction being available. Moreover there is a problem with the hypothesized main clause, since it again does not obey the verb-second condition of German.



Therefore the interpretation is changed to a relative clause, which however cannot appear in isolation. The valency requirements of the verb ‘gräbt’ are satisfied by taking the indefinite pronoun ‘anderen’ as a direct object with the true object (‘eine Grube’) as a (mal-formed) apposition.



Finally, the analysis switches to an interpretation which accepts the second part of the sentence as the main clause and subordinates the first part as a subject clause. The problem with the apposition reading persists.



By interpreting the indefinite pronoun as an ethical dative, the direct object valence is freed for the NP ‘eine Grube’. Although this structure still violates some constraints (e.g. the ethical dative is slightly penalized for being somewhat unusual) a better one cannot be found. Note that the algorithm does not take the shortest possible transformation sequence; in fact, the first analysis could have been transformed directly into the last by only one exchange. Because the algorithm is *greedy*, it chooses a different repair at that point, but it still finds the solution in about three seconds on a 3 GHz Pentium machine.

In contrast to stochastic parsing approaches, a WCDG can be modified in a specifically targeted manner. It therefore provides us with a grammar formalism which is particularly well suited to precisely measure the contributions of different linguistic knowledge sources to the overall parsing quality. In particular it allows us to

1. switch off constraints, i.e. increase the space of acceptable constructions and/or syntactic structures,
2. weaken constraints, by changing the weight in a way that it makes the violation of the constraint condition more easily acceptable,
3. introduce additional dependency labels into the model,
4. remove existing dependency labels from the model
5. reinforce constraints, by removing guards for exceptional cases from them,
6. reinforce constraints, by strengthening their weights or making the constraint non-defeasible in the extreme case, and
7. introducing new constraints, to prohibit certain constructions and/or syntactic structures.

Since for the purpose of our experiments, we start with a fairly broad-coverage grammar of German, from which certain rare phenomena will be removed, options 4 to 7 are most important for us.

4 Robust behaviour under limited coverage

In general, it is not easy to predict the possible outcome of a parsing run when using a grammar with a reduced coverage. Whether a sentence can be analysed at all solely depends on the available alternatives for structuring it. Which structural description it can receive, however, is influenced by the scores resulting from rule applications or constraint violations. Moreover, the transformation-based solution method used for the WCDG-experiments introduces yet another condition: since it is based on a limited heuristics for candidate generation, the grammar must license not only the final parsing result for a sentence, but also all the intermediate transformation steps with a sufficiently high score. This might exclude some structural interpretations from being considered at all if the grammar is not tolerant enough to accommodate highly deviant structures.

Thus, the ability to deal with extragrammatical input in a robust manner is a crucial property if we are going to use a grammar with coverage limitations. Unfortunately, robust behaviour is usually achieved by *extending* instead of *reducing* the coverage of the model and compensating the resulting increase in ambiguity by an appropriately designed scoring scheme together with an optimization procedure.

To deal with these opposing tendencies, it is obviously important to determine which parts of the model need to be relaxed to achieve a sufficient degree of robustness, and which ones can be reinforced to limit the space of alternatives in a sensible way. Excluding phenomena from the grammar which never occur in a corpus should always give an advantage, since this reduces the number of alternatives to consider at each step without forbidding any of the correct ones.

On the other hand, removing support for a construction that is actually needed forces the parser to choose an incorrect solution for at least some part of a sentence, so that a deterioration might occur instead. But even if coverage is reduced below the strictly necessary amount, a net gain in accuracy could occur for two reasons:

1. Leaking: The grammar overgenerates the construction in question, so that forbidding it prevents errors occurring on 'normal' sentences.

2. Focussing: Due to a more restricted search space, the parser is not led astray by rare hypotheses, thus saving processing time which can be used to come closer to the optimum.

4.1 Experiment 1: More with less

In our first experiment, we analysed 10,000 sentences of online newscast texts both with the normal grammar and with the 21 rare phenomena explicitly excluded. As usual for dependency parsers, we measure the parsing quality by computing the structural accuracy (the ratio of correct subordinations to all subordinations) and labelled accuracy (the ratio of all correct subordinations that also bear the correct label to all subordinations). Note that the WCDG parser always establishes exactly one subordination for each word of a sentence, so that no distinction between precision and recall arises. Also, the grammar is written in such a way that even if a necessary phenomenon is removed, the parser will at least find *some* analysis, so that the coverage is always 100%.

As expected, those 'rare' sentences in which at least one of these constructions does actually occur are analyzed less accurately than before: structural and labelled accuracy drop by about 2 percent points (see Table 3). However, the other sentences receive slightly better analyses, and since they are in the great majority, the overall effect is an increase in parsing quality. Note also that the 'rare' sentences appear to be more difficult to analyze in the first place.

Grammar:	Normal	Reduced
Online newscasts		
rare (717)	87.6%/85.2%	85.8%/85.8%
normal (9,283)	91.0%/89.8%	91.4%/90.4%
overall (10,000)	91.0%/89.4%	91.3%/89.7%
NEGRA corpus		
rare (91)	85.5%/83.7%	84.0%/81.4%
normal (909)	91.2%/89.3%	91.5%/89.7%
overall (1,000)	90.5%/88.6%	90.6%/88.7%

Table 3: Structural and labelled accuracy when parsing the same text with reduced coverage.

The net gain in accuracy might be due to plugged leaks (misleading structures that used to be found are rejected in favor of correct structures) or to focussing (structures that were preferred but missed through search errors are now found). A point in case of the latter explanation is the fact that the average runtime decreases by 10% with the reduced grammar. Also, if we consider only those sentences on which the local search originally exceeded the time limit of 500 s and therefore had to be interrupted, the accuracy rises from 85.2%/83.0% to 86.5%/84.4%, i.e. even more pronounced than overall.

4.2 Experiment 2: Stepwise refinement

For comparison with previous work and to investigate corpus-specific effects, we repeated the experiment with the test set of the NEGRA corpus as defined by (Dubey and Keller, 2003). For that purpose the NEGRA annotations were automatically transformed to dependency trees with the freely available tool DEPSY (Daum et al., 2004). Some manual corrections were made to its output to conform to the annotation guidelines of the WCDG of German; altogether, 1% of all words had their regents changed for this purpose.

Table 3 shows that the proportion of sentences with rare phenomena is somewhat higher in the NEGRA sentences, and consequently the net gain in parsing accuracy is smaller; apparently the advantage of reducing the problem size is almost cancelled by the disadvantage of losing necessary coverage.

To test this theory, we then reduced the coverage of the grammar in smaller steps. Since constraints allow us to switch off each of the 21 rare phenomena individually, we can test whether the effects of reducing coverage are merely due to the smaller number of alternatives to consider or whether some constructions affect the parser more than others, if allowed.

We first took the first 3,000 sentences of the NEGRA corpus as a training set and counted how often each construction actually occurs there and in the test set. Table 4 shows that the two parts of the corpus, while different, seem similar enough that statistics obtained

Nr	Phenomenon	Frequency per 1000 on training set	test set
1	*Mittelfeld* extraposition	33.3	13.0
2	ethical dative	16.7	15.0
3	Nominalization	20.3	17.0
4	Vocative	1.0	2.0
5	Parenthetical matrix clause	13.3	23.0
6	verb-first subclause	8.0	3.0
7	Headline phrase	6.7	9.0
8	coordination cluster	1.7	2.0
9	Adverbial pronoun	4.0	2.0
10	*um* omission	1.3	1.0
11	Metagrammatical usage	5.7	2.0
12	Auxiliary flip	2.0	0.0
13	Adjectival subclause	0.0	1.0
14	Suffix drop	1.0	1.0
15	Elliptical genitive	0.0	1.0
16	Adverbial noun	0.0	0.0
17	Verb/particle mismatch	0.0	0.0
18	*Vorfeld* extraposition	0.0	1.0
19	double relative subject	0.0	0.0
20	Relative subject clause	0.3	0.0
21	NP extraposition	0.0	0.0

Table 4: Comparison of training and test set.

on the one could be useful for processing the other. The test set was then parsed again with the coverage successively reduced in several steps: first, all constructions were removed that *never* occur in the training set, then those which occur less than 10 times or 100 times respectively were also removed. We also performed the opposite experiment, first removing support for the least rare phenomena and only then for the really rare ones.

Phenomena removed	structural accuracy	labelled accuracy
none	90.5%	88.6%
$= 0$	90.5%	88.7%
< 10	90.6%	88.8%
< 100	90.7%	88.6%
$>= 100$	90.5%	88.6%
$>= 10$	90.4%	88.5%
> 0	90.5%	88.6%
all	90.6%	88.7%

Table 5: Parsing with coverage reduced stepwise.

Table 5 shows the results of parsing the test set in this way (the first and last lines are repetitions from Table 3). The resulting effects are very small, but they do suggest that removing coverage for the very rare constructions is somewhat more profitable: the first three new experiments tend to yield better accuracy than the original grammar, while in the last three it tends to drop.

4.3 Experiment 3: Plugging known leaks

The previous experiment used only counts from the treebank annotations to determine how rare a phenomenon is supposed to be, but it might also be important how rare the parser actually assumes it to be. The fact that a particular construction never occurs in a corpus does not prevent the parser from using it in its analyses, perhaps more often than another construction that is much more common in the annotations. In other words, we should measure how much each construction actually leaks. To this end, we parsed the training set with the original grammar and grouped all 21 phenomena into three classes:

A: Phenomena that are predicted much more often than they are annotated

B: Phenomena that are predicted roughly the right number of times

C: Phenomena that are predicted less often than annotated (or in fact not at all).

'Much more often' here means 'by a factor of two or more'; constructions which were never predicted *or* annotated at all were grouped into class C.

There are different reasons why a phenomenon might leak more or less. Some constructions depend on particular combinations of word forms in the input; for instance, an auxiliary flip can only be predicted when the finite verb does in fact precede the full verb (phenomenon 12 in Table 1), so that covering it should not change the behaviour of the system much. But most sentences contain more than one noun phrase which the parser might possibly misrepresent as a non-projective extraposition (phenomenon 1). Also, some rare phenomena are dispreferred more than others even when they are allowed. We did not investigate these reasons in detail.

Phenomena removed	structural accuracy	labelled accuracy
none	90.5%	88.6%
A (1,3,4,6–10,13,16,18–21)	90.9%	89.0%
B (2,5,11,12)	90.4%	88.5%
C (14,15,17)	90.4%	88.6%
1–21	90.6%	88.7%

Table 6: Parsing with coverage reduced by increasing leakage.

Table 6 shows an interesting asymmetry: of our 21 constructions, 14 regularly leak into sentences where they have no place, while 4 work more or less as designed. Only 3 are predicted too seldom. This is consistent with our earlier interpretation that most added coverage is in fact unhelpful when judging a parser solely by its empirical accuracy on a corpus.

Accordingly, it is in fact more helpful to judge constructions by their observed tendency to leak than just by their annotated frequency: the first experiment (A) yields the highest accuracy for the newspaper text. Conversely, removing those constructions which actually work largely as intended (B) reduces even the overall accuracy, and not just the accuracy on 'rare' sentences. The third class contains only three very rare phenomena, and removing them from the grammar does not influence parsing very much at all.

Note that this result was obtained although the distribution of the phenomena differs between parser predictions on the training set and the test set; had we classified them according to their behaviour on the test set itself, the class A would have contained only 9 items (of which 7 overlap with the classification actually used).

5 Related work

The fact that leaking is an ubiquitous property of natural language grammars has been noted as early as 80 years ago by (Sapir, 1921). Since no precise definition was given, the notion offers room for interpretation. In general linguistics, leaking is usually understood as the underlying reason for the apparent impossibility to write a grammar which is complete, in the sense that it covers all sentences of a language, while maintaining a precise distinction between correct an incorrect word form sequences (see e.g. (Sampson, forthcoming)). In Computational Linguistics, attention was first drawn to the resulting consequences for obtaining parse trees when it became obvious that all attempts to build wide-coverage grammars led to an increase in output ambiguity, and that even more fine-grained feature-based descriptions were not able solve the problem. Stochastic approaches are usually considered to provide a powerful countermeasure (Manning and Schütze, 1999). However, as (Steedman, 2004) already noted, stochastic models do not address the problem of overgeneration directly.

Disregarding rare phenomena is something that can be achieved in a stochastic framework by putting a threshold on the minimum number of occurrences to be considered. Such an approach is mainly used to either exclude rare phenomena in grammar induction (c.f. (Solsona et al., 2002)) or to prune the search space by adjusting a beam width during parsing itself (Goodman, 1997). The direct use of thresholding techniques at the level of the stochastic model, however, has not been investigated extensively so far. Stochastic models of syntax suffer to such a degree from data sparseness that in effect strong efforts in the opposite direction become necessary: instead of ignoring rare events in the training data, even unseen events are included by smoothing techniques. The only experimental investigation of the impact of rare events we are aware of is (Bod, 2003), where heuristics are explored to constrain the model in the DOP framework by ignoring certain tree fragments. Contrary to the results of our experiments, very few constraints have been found that do not decrease the parse accuracy. In particular, no improvement by disregarding selected observations was possible.

The tradeoff between processing time and output quality which our transformation-based problem solving strategy exhibits, is also a fundamental property of all beam-search procedures. While a limited beam width might cause search errors, widening the beam in order to improve the quality requires investing more computational resources (see e.g. (Collins, 1999)). In contrast to our transformation-based procedure, however, the commonly used Viterbi search is not interruptible and therefore not in a position to really profit from the tradeoff. Thus, focussing as a possibility to increase output quality to our knowledge has never been investigated elsewhere.

6 Conclusions and future work

We have investigated the effect of systematically reducing the coverage of a general grammar of German. By removing support for 21 rare phenomena, the overall parsing accuracy could be improved. We confirmed the initial assumption about the effects that broad coverage has on the parser: while it allows some special sentences to be analysed more accurately, it also causes a slight decrease on the much more numerous normal sentences.

This result shows that at least with respect to this particular grammar, *more* coverage can indeed lead to *less* parsing accuracy. In the first experiment we measured the overall loss through adding coverage where it is not needed as about 0.4% of structural accuracy on newscast text, and 0.1% on NEGRA sentences. This figure can be interpreted as the result of overgenerating or 'leaking' of rare constructions into sentences where they are not wanted.

Although we found that it makes little difference whether to remove support for very rare or for somewhat rare phenomena, judging constructions by how many leaks they actually cause leads to a greater improvement. On the NEGRA test set, removing the 'known troublemakers' leads to a greater increase of in accuracy of 0.4%, reducing the error rate for structural attachment by 4.2%.

Of course, removing rare phenomena is not a viable technique to substantially improve parser accuracy, if only for the simple fact that it does not scale up. However, it confirms that as soon as a certain level of coverage has been reached, robustness, i.e. the ability to deal with unexpected data, is more crucial than coverage itself to achieve high quality results on unrestricted input.

On the other hand, the improvement we obtained is not

very large, compared to the already rather high overall performance of the parser. This may be due to the consistent use of weighted constraints in the original grammar, which slightly disprefer many of the 21 phenomena even when they are allowed, and we assume that the original grammar is already reasonably effective at preventing leaks. This claim might be confirmed by reversing the experiment: if all phenomena were allowed *and* all dispreferences switched off, we would expect even more leaks to occur.

To carry out comparable experiments on generative stochastic models presents us with the difficulty that it would first be necessary to determine which of its parameters are responsible for covering a specific phenomenon, and whether they can be modified as to remove the construction from the coverage without affecting others as well. Even in WCDG it is difficult to quantify how much of the observed improvement results from plugged leaks, and how much from focussing. This could only be done by observing all intermediate steps in the solution algorithm, and counting how many trees that were used as intermediate results or considered as alternatives exhibit each phenomenon.

The most promising result from the last experiment is that it is possible to detect particularly detracting phenomena, which are prime candidates for exclusion, in one part of a corpus and use them on another. This suggests itself to be exploited as a method to automatically adapt a broad-coverage grammar more closely to the characteristics of a particular corpus.

References

Steven Abney. 1996. Statistical Methods and Linguistics. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 1–26. The MIT Press, Cambridge, Massachusetts.

Rens Bod. 2003. Do all fragments count? *Natural Language Engineering*, 9(4):307–323.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proc. 1st Meeting of the North American Chapter of the ACL, NAACL-2000*, Seattle, WA.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, Philadephia, PA.

Michael Daum, Kilian Foth, and Wolfgang Menzel. 2004. Automatic transformation of phrase treebanks to dependency trees. In *Proc. 4th Int. Conf. on Language Resources and Evaluation*, pages 99–106, Lisbon, Portugal.

Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. 1994. XTAG system - A Wide Coverage Grammar for English. In *Proc. 15th Int. Conf. on Computational Linguistics, COLING-1994*, pages 922 – 928, Kyoto, Japan.

Amit Dubey and Frank Keller. 2003. Probabilistic Parsing for German using Sister-Head Dependencies. In *Proc. 41st Annual Meeting of the Association of Computational Linguistics, ACL-2003*, Sapporo, Japan.

Kilian Foth, Michael Daum, and Wolfgang Menzel. 2005. Parsing unrestricted German text with defeasible constraints. In H. Christiansen, P. R. Skadhauge, and J. Villadsen, editors, *Constraint Solving and Language Processing*, volume 3438 of *Lecture Notes in Artificial Intelligence*, pages 88–101, Berlin. Springer-Verlag.

Joshua Goodman. 1997. Global thresholding and multiple-pass parsing. In *Proc. 2nd Int. Conf. on Emprical Methods in NLP, EMNLP-1997*, Boston, MA.

C. Grover, J. Carroll, and E. Briscoe. 1993. The Alvey natural language tools grammar (4th release). Technical Report 284, Computer Laboratory, University of Cambridge.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Natural Language Processing*. MIT Press, Cambridge etc.

Geoffrey Sampson. forthcoming. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*.

Edward Sapir. 1921. *Language: An Introduction to the Study of Speech*. Harcourt Brace, New York.

Ingo Schröder. 2002. *Natural Language Parsing with Graded Constraints*. Ph.D. thesis, Department of Informatics, Hamburg University, Hamburg, Germany.

Roger Argiles Solsona, Eric Fosler-Lussier, Hong-Kwang J. Kuo, Alexandros Potamianos, and Imed Zitouni. 2002. Adaptive language models for spoken ddialogue systems. In *Proc. Int. Conf. on Acoustics, Seech, and Signal Processing, ICASSP-2002*, Orlando, FL.

Mark Steedman. 2004. Wide Coverage Parsing with Combinatory Grammars. Slides of a seminar presentation, Melbourne University, Australia. `http://www.cs.mu.oz.au/research/lt/seminars/steedman.pdf`. Last time visited: 2006-01-06.