Proceedings of the Workshop on Exploring Syntactically Annotated Corpora 14th July 2005

held in conjunction with the Corpus Linguistics 2005 conference
University of Birmingham, 14-17 July 2005

Supported by:

BulTreeBank Project
Bulgarian IST Centre of Competence in 21 Century: BIS-21++

Editors: Kiril Simov, Dimitar Kazakov and Petya Osenova

Birmingham University 2005

Workshop Organisers

Kiril Simov
BulTreeBank Project
Linguistic Modelling Laboratory, IPP,
Bulgarian Academy of Sciences

Dimitar Kazakov
Dept. of Computer Science, University of York

Petya Osenova
BulTreeBank Project
Linguistic Modelling Laboratory, IPP,
Bulgarian Academy of Sciences

Workshop Programme Committee

Table of Contents

Preface

A lot of syntactically annotated corpora have recently been created for various languages. Therefore, the question of applicability and usefulness of such resources seems to have become of great importance. Syntactically annotated corpora can be viewed in two ways. Firstly, they are a reliable base for resource creation, whether for further annotation, such as semantic and discourse annotation, automatic extraction of linguistic knowledge (grammar and lexicon extraction, creation of automatic parsing tools, etc.) or as an environment for testing tools. Secondly, these corpora can be seen as a base for navigation and search playing their role in the development of query languages and support engines.

This Workshop on Exploring Syntactically Annotated Corpora was organised as a joint effort between the Linguistic Modelling Laboratory (LML) at the Institute for Parallel Processing of the Bulgarian Academy of Sciences, and the Artificial Intelligence Group at the Department of Computer Science of the University of York, UK. The workshop is the latest in a series of linguistics workshops and tutorials organised by members of either reseach team, e.g., at the European Summer School in Logic, Language and Information (ESSLLI) (Nancy, France) in 2004 and at the Conference on Corpus Linguistics (CL) (Lancaster, UK) in the previous year. We hope that this workshop will follow step in the fruitful discussions of the Workshop on Shallow Processing of Large Corpora (SProLaC), organised in the framework of the previous CL conference in 2003. We want to thank all reviewers for their support and helpful comments, and all participants for their contributions.

Here is a general overview of the contributors' papers: Paula Chesley and Susanne Salmon-Alt's paper is titled "From the Corpus to the Lexicon: the Example of Data Models for Verb Subcategorization". It focuses on exploring linguistic information in corpora for constructing a rich lexical database for French. Lilja Øvrelid presents some recent results on automatic classification of Norwegian nouns with respect to animacy. Her paper is titled "Animacy Classification based on Morphosyntactic Frequencies". Marina Santini's paper "Building on Syntactic Annotation: Labelling Subordinate Clauses" presents heuristics for labelling subordinate clauses, which are based on syntactic patterns over the result from the parser. Evita Linardaki shows two alternative solutions to the problem of the estimation of the probabilities for syntactic descriptions from different domains. Her paper is under the title "Comparing Constituents of Different Categories in DOP". Aline Villavicencio and Louisa Sadler report on the typology of the agreement patterns in Portuguese through a corpus frequency research. Their paper is titled "Agreement Patterns in Syntactically Annotated Corpora".

Finally, we would like to thank the organisers of CL-05 for providing us with the excellent opportunity for organising this event and for their support in preparing it.

Kiril, Petya, Mitko

14 July 2005
Birmingham

From the corpus to the lexicon: the example of data models for verb subcategorization

PAULA CHESLEY
Linguistics Department, University at Buffalo, USA & ATILF-CNRS, France
pchesley@buffalo.edu

SUSANNE SALMON-ALT
ATILF-CNRS, France
susanne.alt@loria.fr

Abstract

This paper describes the integration of corpus-based syntactic subcategorization frames and correlated semantic information into a large-scale, cross-theoretically informed lexical database for French (Romary et al. (2004)). This database is the first to implement the Lexical Markup Framework (LMF), an international initiative towards ISO standards for lexical databases (ISO TC 37/SC 4). The subcategorization frames have been acquired via a dependency-based parser (Bick (2003)), whose verb lexicon is currently incomplete with respect to subcategorization frames. Therefore, we have implemented probabilistic filtering as a post-parsing treatment using the binomial distribution. Building on our discussion of what semantic information, e.g., participant roles, to include in the database, we describe how we plan to exploit our findings on subcategorization frames to derive this information via unsupervised learning techniques.

1 Introduction

This paper describes the integration of corpus-based syntactic subcategorization frames and correlated semantic information into a large-scale, cross-theoretically informed lexical database for French (Romary et al. (2004)). The Morphalou database is freely downloadable[1] and is the first to implement the Lexical Markup Framework (LMF), an international initiative towards ISO standards for lexical databases (ISO TC 37/SC 4). We thus discuss how the recommended LMF data structures for syntactic and semantic information guide our acquisition of subcategorization frames of verbs from annotated corpora, as well as the representation of the syntax-semantics interface in wide-coverage lexical databases. Because automatic subcategorization extraction for French is a nascent field, we have chosen to initially concentrate on extracting subcategorization information of verbs. Subsequently, we aim to extend this research to other parts of speech that also exhibit subcategorization phenomena.

At present there is no lexical database for French that encodes lexical syntactic and semantic information to the extent that the current research plans to do so. For example, the French component of EuroWordNet does not contain any subcategorization frame or argument structure information. Furthermore, establishing a standard lexical database format ensures that information in this format will remain functional and exploitable for many years to come. Thus, each component of the current undertaking is an essential step for Natural Language Processing (NLP) for French.

As Briscoe and Carroll (1997) note in a study on English, incorrect syntactic subcategorization information is responsible for approximately half of parsing errors like incorrect prepositional attachments. Assuming similar figures for French, any French parser would benefit from subcategorization information that the Morphalou database will contain. In addition, monolingual and bilingual dictionaries note subcategorization information. The manual alternative to automatic extraction of subcategorization frames for dictionaries is of course expensive, time-consuming, and potentially difficult to re-use.

[1] More information about the database can be found at http://www.atilf.fr/morphalou.

Additionally, Gildea and Jurafsky (2002) note that domain-specific semantic information is employed in spoken dialogue and information extraction systems, but that there is yet a lack for general semantic information such as that of Fillmore's (1976) semantic frames. Having such related information, the authors note, would allow verbs of a common frame, such as *send* and *receive* from the TRANSFER frame, to share the same semantic roles, thus aiding in a question-answering system in which one verb is in the question, while another is in the response. Furthermore, Nasr (2004) notes that, inter alia, lack of argument structure information in an annotated learning corpus for French could constitute a reason for which his dependency parser did not fare as well for French as it did on the English test corpus containing such information.

Following existing initiatives in the modeling of syntactic and semantic lexical knowledge (Genelex, ISLE/MILE, etc.), the integration of subcategorization frames into the LMF data model occurs at several levels which are both syntactic and semantic in nature. On the syntactic side, one central data structure characterizes a lexical entry, that of a set of syntactic constructions. This construction set corresponds to a set of frames observable in a corpus for a given entry with a given sense. Each syntactic construction is further described by a set of syntactic positions, i.e., the syntactic category (nominal phrase, subordinate clause, etc.) and syntactic function (subject, direct object, etc.) of each element subcategorized for by the verb. Further information about the LMF data model is found in section 2.

Since semantic information is not directly observable in corpora for French, we aim to use corpus-observed subcategorization information to infer semantic knowledge to be incorporated in the lexical database. Our experiment is based on a corpus we have created from Frantext, an online literary French database. The subcategorization frames have been acquired via a dependency-based parser (Bick (2003)), whose verb lexicon is currently incomplete with respect to subcategorization frames. Therefore, we have implemented probabilistic filtering as a post-parsing treatment using the binomial distribution. This sort of dual treatment constitutes a technique shown to be successful in subcategorization frame filtering for English (Brent (1992), Brent (1993), Manning (1993), Briscoe and Carroll (1997)). Our study differs from that of Bourigault and Frérot (2005), who are undertaking research in subcategorization of prepositional phrases (PPs). That is, in addition to PPs, the present work seeks to extract subordinate clauses, the impersonal *il* subject, and direct and indirect objects in complete frames. Section 3 details the subcategorization frame extraction process.

Determining a verb's semantic traits to be integrated into the Morphalou database relies on work in linking in theoretical linguistics. Once decided upon, we advance the hypothesis that semantic information can be uncovered via unsupervised learning of observable linking phenomena and surface cues subcategorization frames in corpora. However, even from correct subcategorization frames, the proper number of semantic arguments for the verb is difficult to obtain, since some participant roles are optional (Koenig et al. (2003)). We can, nevertheless, infer some information about participant roles, since for at least some verbs and some subcategorization frames there is a direct correlation to a semantic argument. For example, the preposition *vers*, 'toward' in a PP complement, maps to a semantic *locative* or *goal* argument. However, not all surface cues will yield dependable semantic information. The semantic aspects of the database are discussed in section 4.

2 LMF, an international standard for lexical databases

Lexical structures can classically be considered according to the way they organize the relation between words and senses. On the semasiological view, senses are considered as subdivisions of the lexical entry, whereas on the onomasiological view, words are considered as ways of expressing concepts. Of these views, the former allows an exhaustive survey of lexical content for a given language.

In particular, it corresponds to the basis for any classical editorial, or print, dictionary, and also underlies, at least implicitly, most existing NLP lexicons. From a theoretical perspective, the internal structure of a lexical entry can be configured through different layers. In a two-layered approach, the `/form/` and `/sense/` layers are anchored to the Saussurian definition of a linguistic sign and are related to the basic notions of *signifier*, the sound pattern of a lexical entry, and *signified*, the corresponding concept. The syntactic behavior of the lexical unit is thus systematically subjacent to its semantic description. This notion is currently being implemented in the LMF and is being developed in the ISO TC 37/SC 4 as a future standard for the representation of lexical resources (Francopoulo et al. (2004)). Accordingly, the LMF core model is organized as a hierarchical structure built upon the following components:

- the `/lexicalDatabase/` component, which gathers all information related to a given lexicon;
- the `/globalInformation/` component collecting metadata such as version number, contributors, updates made, etc.;
- a `/lexicalEntry/` component, which corresponds to the elementary lexical unit in a lexical database;
- a `/form/` component providing access to surface properties, i.e., phonological and graphical realizations as well as grammatical properties such as inflectional features;
- one or more `/sense/` components, which currently organize the lexical entry. These components can be repeated, in the case of homonymy, and further divided into sub-senses in the case of polysemy.

Furthermore, following general principles of the linguistic annotation scheme design stated in Ide and Romary (2003), the LMF provides a mechanism for combining the components of the basic data model with elementary descriptors, or data categories. Data categories reflect basic morphosyntactic concepts (e.g. `/partOfSpeech/`, `/grammaticalNumber/`, `/grammaticalCase/`, etc.). They are stored and managed independently from the hierarchical structure of the data model. Proceeding in this way allows for recording language-specific properties independently of structural properties of the linguistic layers to be described. For instance, the data category `/grammaticalGender/` holds two values for French, `/masculine/` and `/feminine/`, and three values for German, `/masculine/`, `/feminine/` and `/neuter/`. In order to share data categories within the community, the ISO/TC 37 deploys an online data-category registry[2] for use in conjunction with the other standardization activities. The future LMF standard as such does not aim to provide a specific list of data categories to be used for lexical descriptions. Doing so would be far too complex, given the potential variety of applications. It is thus expected that implementers will systematically refer to the ISO/TC 37 data category registry to find the proper descriptive background for their individual needs.

Finally, the LMF provides mechanisms to translate the combination of the core model and data categories into an isomorphic XML pivot structure. The implementers might then chose to express their own combination of a core model and data categories in an LMF-XML "dialect". For example, it is possible to implement a given data category such as `/grammaticalGender/` as an XML element rather than an attribute, or by renaming it as `/gen/`, `/gender/` or `/genre/`. Crucially, such a proprietary XML dialect must be able to be mapped unambiguously to the LMF-compatible XML pivot structure in order to ensure proper standardization.

[2]This registry is accessible at `http://syntax.inist.fr`.

2.1 Extending the LMF to syntax and semantics

In its current state, the syntactic extension of LMF essentially covers syntactic realizations of argument structures for entries with predicative senses, especially verbs. The researchers involved have not yet come to a consensus on the ensemble of components to be used. In the present work, our description thus proceeds from the concrete LMF structures, i.e., a model for data structures directly observable in a corpus, to the more abstract. The most concrete data structures in the syntactic component are at the level of the syntactic dependent, the syntactic realization of semantic argument, the data category for which is `/syntacticArgument/`. A syntactic dependent is minimally described by the following data categories:

- `/syntacticFunction/`, having basic values such as `/subject/`, `/directObject/` and `/prepositionalObject/` which might be able to be refined with user defined data categories for language-specific phenomena;
- `/syntacticConstituent/`, describing the syntactic category of the argument, e.g. `/nounPhrase/`, `/prepositionalPhrase/`, `/subordinateClause/`, etc.;
- `/syntacticIntroducer/`, allowing the user to record, in case of prepositional phrases or subordinate clauses, the preposition or the complementizer. This data category can of course be extended to languages that make use of other introducers like postpositions for Korean.

In addition, each syntactic dependent has a /`semanticRestriction`/ data category. This data category can contain participant roles as values as well as other lexical semantic information. More data categories for the syntactic dependent level may be added according to further research.

The LMF allows for a recursive description of subordinate clauses in terms of a set of syntactic dependents, thus providing a simple way of encoding various morphosyntactic constraints on subordinates such as mood, tense and co-indexation of subject or object. It is also possible to add examples or occurrence frequencies at various levels of granularity. This can prove useful for illustrating a particular syntactic dependent with a corpus example or to count the occurrences of a particular realization of a syntactic dependent.

Building upward, the first level of abstraction is the syntactic construction, with the data category `/syntacticConstruction/`, which represents a subcategorization frame. We define a subcategorization frame as a set of syntactic dependents, realized simultaneously by a predicative lexical entry. The introduction of this component corresponds to the need of an anchoring point for lexical information about the whole construction, e.g., the auxiliary verb for past participle forms, or constraints on ordering and/or the simultaneous realisation of different syntactic arguments. More fundamentally, syntactic constructions are the basic data structures on which syntactic alternations and transformations are effectuated. Depending on the degree of extensionality of the lexicon, the user might decide to explicitly encode the entire range of surface syntactic constructions (including, for example, passive constructions), or to encode only canonical constructions (`/canonicalConstruction/`) to be associated with grammatical rules if the lexicon is to be used in conjection with a parser.

A further abstraction is the grouping of those abstract constructions into classes sharing the same syntactic behavior with respect to alternations and transformations. In the LMF, this component is referred to as a verb's lexical class and represents the crucial point of the syntax-semantics interface in the model. It can be understood as a class of verbs similar to Levin's (1993) seminal work for English, or a table number from Maurice Gross' (1975) tables denoting the syntax-semantics interface for French verbs. This and other aspects of the LMF architecture are given in the example of a lexical entry in appendix B.

Having established the outline of the LMF in its current state, we turn now to the practical concerns of incorporating lexical information for French into the Morphalou database.

3 Subcategorization frame extraction

Extensive work in subcategorization frame extraction for French has been carried out by Bourigault and Frérot (2005). These experiments have been effectuated on large-scale corpora of approximately 200 million words. The focal point of these experiments is subcategorization of PPs; thus, subordinate clauses, impersonal subjects, and nominal and adjectival attributes are not discussed in detail. In addition, PPs are treated independently of other syntactic dependents – in fact, all syntactic dependents are treated independently of each other. In effect, this research does not obligatorily correlate a subcategorized PP to an attested frame: if the PP is dependent on another dependent that is absent, the PP might be misanalyzed as a dependent for which the verb subcategorizes rather than a modifier. An example of this phenomenon is given in (1).

(1) a. Jean a reçu un message de (la part de) la sécrétaire.
"Jean received a message from the secretary."

b. Jean a reçu de la confiture (pour son anniversaire).
"Jean received jam (for his birthday)."

In (1a), *de* is a preposition introducing a subcategorized PP, while in (1b) it is an indefinite article in the direct object. If we do not take into account other syntactic dependents – in this case, the direct object of (1a), we risk misparsing the direct object of (1b) as a PP. However, considering syntactic dependents independently of each other might prove sufficient most of the time. In addition, examining co-occurrences of syntactic dependents could induce errors due to subcategorization frame information that is too fine-grained. We simply felt it more judicious to err initially on the side of caution than to be potentially obliged to change our frame extraction methodology.

The present work on subcategorization frame extraction is based on a corpus of 115 verbs that we created from the online literary database Frantext. Our corpus comes from various genres between the years 1850 and 2000, such as treaties and novels, and excludes theatre and poetry, since these genres can yield statistically higher percentages of non-canonical subcategorization realizations than prose. With the query tool that Frantext provides, we have also excluded most occurrences of the causative construction, since it can change the argument structure, and thus the subcategorization frames, of a verb. Given these restrictions, we randomly chose 200 occurrences each of 115 verbs in the TSNLP for French to be parsed. A list of these verbs is given in appendix A. We used the VISL parser (Bick (2003)), a dependency-based parser whose lexicon is partially complete with respect to subcategorization frames[3]. Although in the parser analysis some dependent positions are more reliable than others we have chosen to weight all dependent positions equally and to let the filtering choose the correct subcategorization frames, since presumably filtering will also correctly label those which the parser does. We thus use the parser almost as a chunker that divides a sentence into phrases.

After the parsing stage, we effectuate a probabilistic filtering treatment that makes use of the binomial distribution. Dual treatments using this filtering method have proven extremely successful for subcategorization frame extraction for English (Brent (1993), Manning (1993), Briscoe and Carroll (1997)); for example, precision rates for Brent (1993) vary from 96% to 100% according to the frame. Let a *cue* be an initial frame we receive from the parser, without knowing whether it is indeed a frame.

[3]For a demonstration of the state of the art of the parser the reader may consult `http://visl.hum.sdu.dk/visl/en/parsing/automatic/trees.php`.

The binomial distribution in this application examines the difference between the number of times a particular cue occurs with a given verb and the number of total times the latter appears in the corpus. The greater this difference, the less likely it is that the cue is an actual frame. Let m be the total number of occurrences of a verb in the corpus, n be the number of co-occurrences of the verb with the cue, and B_f the estimated upper bound that the verb that does not subcategorize for the frame f appears nevertheless with f.We make the null hypothesis that the verb does not subcategorize for the cue. The upper bound on the probability that the hypothesis is false given all cues is the following (Manning (1993)):

$$\sum_{i=n}^{m} \frac{m!}{i!(m-i)!} {B_f}^i (1 - B_f)^{m-i}$$

Typical confidence levels are empirically set between .02 and .05, below which the cue is considered an actual frame. In the present work we have set the confidence level at .02.

Note that the binomial distribution supposes a known rate B_f, which is in effect the error rate for each cue. Since Manning (1993) examines 19 frames, he establishes this rate empirically for each of them. Brent (1994) details a method to establish the rate B_f automatically which we have adopted in the current work. In brief, this method consists in examining every occurrence of a frame with every verb above a certain number of occurrences, which we have currently fixed at 50. From these occurrences we construct a histogram based on the number of co-occurrences of cues and the verbs with a sufficient amount of corpus attestations. We look for a binomial distribution toward the lower end of the histogram that signals the false cues f. The average in this distribution is a proper estimation of the rate of false cues B_f. We refer the interested reader to Brent (1994) for an in-depth discussion of the method of finding B_f.

3.1 Evaluation

Once definitive results of this endeavor are established, we can seek to augment the Morphalou database with semantic information that is not only in accordance with theoretical linguistics but also easily exploitable for other NLP applications. We plan to evaluate subcategorization frames of verb types manually against a gold standard and the completed portion of the parser lexicon. Currently we have initial results and are examining what resources to use as a gold standard for a large-scale evaluation of our subcategorization frame extraction. Here are certain results, which we note in contrast to the initial parser results:

- **diriger**, 'to direct'. The parser does not include the frame `Sub V DO PP[`$_{vers}$ `'toward']`, as a frame for this verb, although the Collins Robert English-French dictionary includes it as such.
- **donner**, 'to give'. Our work indicates that `Sub V DO PP[`$_{\grave{a}}$ `'to']` is a frame for this verb. This is the standard ditransitive frame and the parser lexicon includes this frame.
- **courir**, 'to run'. Our experiment supposes the subcategorization frame `Il`$_{imp}$ `V PP[`$_{\grave{a}}$`]`. This frame does not exist in the parser lexicon, although the *Trésor de la langue française informatisé* indicates that it is indeed a subcategorization frame for this verb.
- **arriver**, 'to arrive'. Our experiment supposes the subcategorization frame `Il`$_{imp}$ `V PP[`$_{de}$ `'from']`. In fact, this frame does not exist. A frame with the same surface elements does however exist: `Il`$_{i}mp$ `V CMP[`$_{de}$ `(to)]` ("Il arrive de pleuvoir en été", "It can happen that it rains in summer"). This error could be due to an error in the parser. However, in our informal survey of the results, it appears that frames with the impersonal subject appear more often as part

of subcategorization elements than they should be. We might have to lessen our confidence level for this construction.

Currently bilingual dictionaries appear to have the most explicit subcategorization information. The *Trésor de la langue française informatisé* also appears to have a good amount of subcategorization information, and thus both of these resources could serve as gold standards in our evaluation process.

4 Inferring semantic information from subcategorization frames

Gildea and Jurafsky (2002) use the hand-labeled FrameNet database to build a classifier to discern 18 semantic roles, many of which are included in theoretical research on participant roles. This supervised learning technique is currently unavailable to us, as no hand-annotated corpus of semantic information currently exists for French. There is however large-coverage hand-annotated semantic information available in Gross (1975), in the form of multiple tables, albeit somewhat limited in nature. The traits *locative*, *proposition*, and *human* are consistently given in the tables. Gardent et al. (2005) are currently undertaking research to determine whether the tables are feasibly exploitable in their current format. However, subcategorization extraction must first be ensured in order to guarantee the proper linking bewteen semantic arguments and their syntactic realizations. After discussing what semantic information should be included in the Morphalou database, we address another possibility for obtaining this information, that of combining work in theoretical linguistics on linking and unsupervised learning techniques based on distributions of surface phenomena in the syntactic dependents of verbs in our corpus.

4.1 What semantic information and why?

The current lack of consensus as to what semantic information should be included in the Morphalou database reflects a similar dilemma in theoretical work in lexical semantics and linking. The issue of how the information encoded in a lexical item maps to a surface realization is no trivial issue and cannot be discussed here in great depth. However, as Koenig et al. (2003) note, most scholars are in agreement that "the syntactic structure of many sentences is mostly or entirely determined by the information about situation participants in lexical entries of verbs" (p. 69) (cf. Koenig and Davis (2001)). We thus contend that there is an essential step between a lexical entry and its syntactic realization that contains lexical semantic information; i.e., information about participant roles, that ought to be incorporated into a lexical database.

One way in which to achieve this mapping between semantic arguments onto syntactic dependents is to introduce a set of participant roles, such as *agent*, *patient*, etc. However, Koenig and Davis (2001) note that a principal drawback of this approach is that it cannot itself determine the number and types of participants required. These decisions must be left to linguists, whose opinions could well vary on the matter. Despite this shortcoming, we feel this approach, or variants of it based on current work on argument structure and linking, merit to be examined as a possible way in which to encode semantic information in the Morphalou database.

As much as possible, we would like the Morphalou database to respect a balance between the following concerns in regards to lexical semantic phenomena:

1. Intuitive conceptualization for non-linguists;
2. Linguistic accuracy;
3. A theory-neutral linguistic account.

It is worth noting that points 1 and 2 can at times appear as conflicting goals. For example, the semantic trait *human* is most likely more intuitive to non-linguists than the participant roles of *agent*, *experiencer*, *patient*, or *beneficiary* which a human can fulfill. However, the semantic trait *human* cannot be considered a participant role, while *locative* and *proposition* can be thought of as such. Participant role information should be abstract in nature; i.e., a term used to describe semantic information should not denote an entity existing in the world but rather a linguistic concept. Participant roles seem a more accurate description in light of productive phenomena such as metaphor and metonymy[4]. Additionally, participant roles are implemented as lexical semantic information in FrameNet.

4.2 Automatically acquiring semantic information

As opposed to the manually annotated resources with semantic information described in the tables of Gross (1975), the method we outline for obtaining participant roles has the advantage of being directly exploitable as soon as the evaluation of our subcategorization work is carried out. Clearly manually developed resources constitute a useful gold standard that should be exploited, but recall and precision rates of automatic extraction of subcategorization frames from them have yet to be established. Additionally, the tables do not employ all participant roles, nor do they make the distinction between linguistic concepts and concrete denotations in their semantic information; recall from the previous section that one semantic trait they employ is *human*.

A consensus concerning the ideal number of participant roles to distinguish has not yet been reached in theoretical linguistics. Therefore, we offer the following list as a first proposition of realistic participant roles to be automatically extracted given our corpus of subcategorization frames:

- agent;
- patient;
- location[5];
- instrument;
- beneficiary;
- experiencer;
- proposition.

Certain participant roles will be easier to extract than others. We thus sketch the surface cues and linking distributions which will aid us in extracting the participant roles before discussing possible methods for extracting this information.

In French the *agent* role rarely occurs as a direct or indirect object, barring of course the causative construction. Additionally, functional linguistics indicates that the *agent* role exhibits a strong preference for a syntactic realization of subject (Gildea and Jurafsky (2002)), although this is not always the case. What's more, agents and experiencers tend to be humans, and the distribution of humans in syntactic dependent realizations can be uncovered in using named entities. These facts and other distributional

[4]Markert and Nissim (2003) note that of 1,000 country names examined manually in the BNC, between 171 and 186 of country names constitute metonymical readings (17.1 - 18.6%), as opposed to 737 literal readings (73.7%). This percentage of metonomical uses is clearly significant for country names. A study confounding all named entities might yield similar results. However, it is not sure whether this percentage would still remain significant if all syntactic realizations of argument structure elements are examined.

[5]It remains to be seen if the *location* participant role can be further subdivided into the roles *goal*, *source*, and *destination*.

properties of agents, experiencers, and patients in French must be examined in further detail in order to extract these particpant roles on the basis of syntactic realizations of a verb's semantic arguments.

For the participant roles *beneficiary*, *location* and *proposition*, surface cues such as named entities, prepositions, and clitic realizations aid more in distinguishing these participant roles than for the *agent*, *experiencer*, and *patient* roles. As mentioned in section 1, certain prepositions such as *vers*, 'toward', only map to one participant role, that of *location*[6], when they represent heads of subcategorized PPs. Prepositions like *à* and *de*, respectively 'to' and 'from' or 'of', are far too common and do not map to a particular participant role, but many other prepositions, such as *chez*, 'at the house or establishment of' and *derrière*, 'behind', also demonstrate a direct mapping to a participant role. It is worth noting that these prepositions can also take a metaphorical, non-spatial sense. However, if the participant role changes due to this metaphorical usage, it does not seem probable that verbs subcategorize for these metaphorical senses of these prepositions, while they can and do for the concrete, spatial senses. Similarly, the *instrument* participant role in French is perhaps most often seen with a subcategorization realization of a PP headed by the preposition *avec*, 'with', and propositions are often introduced by the complementizers *de* and *que*, 'that'.

The object indirect in French is most often realized with the preposition *à*. As previously mentioned, this preposition cannot be directly mapped to a particular participant role. However, the indirect object can be cliticized into the unambiguous indirect object clitics *lui* and *leur* (some indirect object clitics also have the same form as direct object clitics). Of the seven participant roles noted above, the indirect object most often informs us about the *beneficiary* participant role.

After surface cues and linking distributions have been established, we can begin the bootstrapping process. Since this research is yet in its elementary stages, we present a simple sketch of how this bootstrapping could take place. Surface cues seem a promising direction. For example, we can say for every verb that subcategorizes for a preposition that maps only to a locative participant role, that the verb takes a locative participant role. The same might well true for a verb that subcategorizes for the preposition *avec*, 'with', and the instrument participant role. In addition, co-occurrence rates of unambiguous indirect object clitics and the verb can be examined, perhaps even in using the binomial distribution with a different confidence level than what we use in the current work, to see whether the verb takes the *beneficiary* partipant role. If we assume that the percentage of realizations of agent, experiencer, and patient roles realized as having the semantic trait *human* vary, a sample set of verbs with known argument structures can be examined for co-occurrence data of realizations with this trait via named-entity recognition. The verbs with similar co-occurrence data could be attributed the same participant roles as the verb with which it shares the most similar co-occurrence rate of *human* realizations.

5 Conclusions and perspectives

This work demonstrates the current state of the LMF structure and the content of the Morphalou database. Crucially, the theoretical work on the structure of the LMF is independent from our work on the content of the Morphalou database for French and will be able to be used for any language. We discuss how theoretical work in linguistics and lexicology influences our choices of structure and data categories. We also show the flexibility of the formats in regards to cross-linguistic diversity and mappings of other formats to the LMF. For example, users of the LMF can choose to use a subset of the data categories proposed in section 2.

The content of the Morphalou database is both syntactic and semantic, and we have detailed not only

[6]This participant role might be confounded with the more abstract role of *goal*.

how we are automatically extracting subcategorization frames for our test corpus of 115 verbs, but also how we plan to derive semantic information, i.e., participant roles of verbs, in the database. This content is similar to the semantic and syntactic information available in the English FrameNet, and we have illustrated our reasoning for selecting this semantic information in section 4.1.

After the evaluation of our subcategorization frame extraction, we principally plan to exploit these frames to derive semantic participant roles to be incorporated in the database. In so doing, we must examine the possibility of more surface cues that can aid in the detection of participant roles, as well as the possibility that current surface cues proposed may be erroneous. In the previous section, we propose that the unambiguous indirect object clitics *lui* and *leur* aid in determining beneficiary roles. However, we can think of at least one example, a psychological verb with the impersonal *il* subject, in which these clitics represent an experiencer rather than a beneficiary:

(2) Il lui plaît de voir sa cousine.
"It pleases him-IO to see his cousin."

Constructions that are contrary to our hypotheses must be examined for inherent patterns as well as frequency rates and productivity. We could also institute a potential default linking system for each frame, e.g., subjects could map to the partipipant role *agent*, according to frequency data of participant roles in a small sample corpus.

References

[Bick2003] E. Bick. 2003. A CG & PSG Hybrid Approach to Automatic Corpus Annotation. pages 1–12. Corpus Linguistics, Lancaster.

[Bourigault and Frérot2005] D. Bourigault and C. Frérot. 2005. Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles.

[Brent1992] M. Brent. 1992. Robust Acquisition of Subcategorization Frames from Unrestricted Text: Unsupervised Learning with Syntactic Knowledge. Master's thesis, Johns Hopkins University, Baltimore, MD.

[Brent1993] M. Brent. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243–262.

[Brent1994] M. Brent, 1994. *Surface Cues and Robust Inference as a Basis for the early Acquisition of Subcategorization Frames*, pages 433–470. MIT Press, Cambridge.

[Briscoe and Carroll1997] T. Briscoe and J. Carroll. 1997. Automatic Extraction of a Subcategorization from Corpora. pages 356–363. Proceedings of the 5th ACL Conference on Applied Natural Language Processing.

[Fillmore1976] C. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280:20–32.

[Francopoulo et al.2004] G. Francopoulo, M. George, and M. Pet. 2004. Data categories in lexical markup framework or how to lighten a model. LREC Workshop "A registry of linguistic data categories within an integrated language resources repository area".

[Gardent et al.2005] C Gardent, B. Guillaume, G. Perrier, and I. Falk. 2005. Maurice gross' grammar lexicon and natural language processing. Proceedings of the 2nd Language and Technology Conference.

[Gildea and Jurafsky2002] D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:3:245–288.

[Gross1975] M. Gross. 1975. *Méthodes en syntaxe: régime des constructions complétives.* Hermann, Paris.

[Ide and Romary2003] N. Ide and L. Romary, 2003. *Encoding Syntactic Annotation*, pages 281–96. Kluwer, Dordrecht.

[Koenig and Davis2001] J.-P. Koenig and A. Davis. 2001. Sublexical modality and the structure of lexical semantic representations. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 24:71–124.

[Koenig et al.2003] J.-P. Koenig, G. Mauner, and B. Bienvenue. 2003. Arguments for Adjuncts. *Cognition*, 89:67–103.

[Levin1993] B. Levin. 1993. *English Verb Classes and Alternations*. Chicago UP.

[Manning1993] C. Manning. 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. pages 235–242. Proceedings of the 31st ACL.

[Markert and Nissim2003] K. Markert and M. Nissim. 2003. Corpus-based metonymy analysis. *Metaphor and Symbol*, 18:3:245–288.

[Nasr2004] A. Nasr. 2004. Analyse syntaxique probabiliste pour grammaires de dépendances extraites automatiquement. Habilitation à diriger des recherches, Université Paris 7.

[Romary et al.2004] L. Romary, G. Francopoulo, and S. Salmon-Alt. 2004. Standards going concrete: from LMF to Morphalou. COLING workshop.

A The 115 verbs in the corpus

aborder	croire	lire	regretter
accepter	croître	livrer	représenter
acheter	decider	maintenir	requérir
agir	démarrer	manger	réserver
aider	devenir	marcher	restaurer
aimer	devoir	marier	rester
aller	dire	mentir	rêver
apercevoir	diriger	mettre	savoir
apparaître	diviser	montrer	séparer
appeler	donner	offrir	signer
apprendre	dormir	ouvrer	sommer
arriver	durer	ouvrir	sortir
asseoir	écrire	paraître	sucrer
avertir	entendre	parler	suer
avoir	entreprendre	participer	suffire
avouer	entrer	partir	suivre
boire	espérer	passer	supposer
causer	étayer	penser	taire
cesser	être	permettre	terminer
combattre	exceller	persuader	tomber
commencer	faillir	plaire	toucher
comparer	faire	pleuvoir	transférer
comprendre	falloir	prendre	travailler
connaître	fontionner	présenter	trouver
constituer	hésiter	prononcer	venir
contrer	indiquer	proposer	vivre
convaincre	intéresser	provoquer	voir
courir	interroger	raconter	vouloir
craindre	laisser	recevoir	

B An example of a lexical entry in the Morphalou database

This format is the current implementation of the LMF. Note that the semantic information in the /`semanticRestriction`/ data category is not in conjunction with what is noted in section 4. This discrepancy is due to the fact that as of yet the /`semanticRestriction`/ information is that contained in the manually annotated tables of Gross (1975). In addition, researchers are still deciding upon what information to use in regards to the /`lexicalClass`/ data category. We limit ourselves to one example subcategorization frame of an entry due to space constraints.

```
<lexicalEntry id="" lemma="alarmer">
    <grammaticalCategory>verb</grammaticalCategory>
       <sense id="1" glose="to trouble" example="Que max parte ennuie Ida"
        source="LADL_table_4">
            <constructionSet source="LADL_table_4">
                <syntacticConstruction exampleConstruction="Max alarme Paul."
                gloss="N0=Nhum N0_V_N1">
                    <syntacticArgument id="a0" canonicalArgument="N0">
                        <syntacticFunction>subject</syntacticFunction>
                        <syntacticCategory>nounPhrase</syntacticCategory>
                        <semanticRestriction>human</semanticRestriction>
                        <semanticRestriction>intentional</semanticRestriction>
                    </syntacticArgument>
                    <syntacticArgument id="a1" canonicalArgument="N1">
                        <syntacticFunction>directObject</syntacticFunction>
                        <syntacticCategory>nounPhrase</syntacticCategory>
                        <semanticRestriction>human</semanticRestriction>
                    </syntacticArgument>
                </syntacticConstruction>
            </constructionSet>
       </sense>
</lexicalEntry>
```

Comparing constituents of different categories in DOP

Evita Linardaki
Dept. of Language & Linguistics
University of Essex
elinaro@essex.ac.uk

Abstract

Syntactically annotated corpora constitute a very popular tool in NLP research. If exploited adequately, they can prove very powerful for training as well as testing grammars. An extension to this is Data Oriented Parsing (DOP), a stochastic NLP model well known for using fragments of a syntactically annotated corpus as a grammar rather than for training some grammar. After parsing, some statistical method is used to resolve ambiguity problems. The issue of the probabilistic algorithm employed in DOP has been the focus of several discussions in the literature, starting from Bod (1992, 1995). In the first few formal instantiations of DOP, a simple counter was used to register frequencies of subtrees, which were subsequently used to calculate the probability of a certain derivation or parse tree. Bonnema et al. (1999) and Bonnema and Scha (2003) proposed an alternative defining the probability of a subtree in terms of its complexity and the relative frequency of the tree it originated from with respect to the sample space of all initial trees with the same root.

Neither of the two models, however, takes into account the fact that before initiating the derivation process, no information regarding the potential root of the constituent to be derived is known. This paper seeks to report on the problems that arise from not taking into account our ignorance about the category of the constituent to be derived. Two alternative solutions are put forward. The first one suggests the incorporation of a pseudo start symbol in the design of the grammar and the second redefining the probability of a derivation in order to reflect the above mentioned observation of known vs unknown information before each derivation step. The two approaches are compared and contrasted and the latter is proposed on the basis of being both computationally more efficient and statistically better justified.

1 Introduction

Data Oriented Parsing (DOP) is a Tree Substitution Grammar (TSG) formalism, different to other TSG formalisms in that it views a decomposed corpus (i.e. a treebank) as a grammar. Moreover, it is a Stochastic TSG (STSG) because using the frequency distribution of subtrees in the treebank, it associates with every tree some probability that indicates the likelihood of that tree being generated by the grammar. Two different probability models have been the center of attention in the DOP literature so far. The first one, proposed by Bod (1992, 1995) and described in more detail in section 3, can be found in most traditional versions of DOP, like DOP1. This model makes use of a simple frequency counter to register frequencies of subtrees and defines the substitution probability of a subtree α as its relative frequency with respect to the number of subtrees with the same root node label, $r(\alpha)$. Johnson (2002) showed this probability model to be biased and inconsistent. The bias was in favour of large corpus trees to which a disproportionate amount of the overall probability mass was assigned. Bonnema et al. (1999) and Bonnema and Scha (2003) proposed an alternative, which will be presented in section 4, defining the substitution probability of a subtree in terms of its complexity and its relative frequency with respect to the sample space of all initial trees rooted at $r(\alpha)$. This model efficiently deals with the problem of assigning disproportionate amounts of the overall probability mass to the subtrees of large corpus trees.

Neither probability model, however, takes into account the fact that the information known before each step of the derivation process differs. On the contrary, the probabilities associated with the subtrees are conditional, presupposing that the root node of a subtree is known before the relevant derivation

step. This, of course, is the case for all subtrees taking part in the derivation process except for the first one. Considering the conditional probability associated with each tree, in this case, causes the probability of the whole derivation to be conditioned upon some predefined root category. Section 5 will show that, as a result, parse tree probabilities can be sensibly interpreted only within the limits of some specific root identifiable sample space, thus disallowing the comparison of trees that belong to different categories. Section 6 will demonstrate one way of solving the problem by incorporating a pseudo start symbol in the design of the grammar. In section 7 we will present an alternative solution based on redefining the probability of a derivation in order to reflect the above mentioned observation of known vs unknown information before each derivation step. The paper concludes in section 8 by a discussion of how our suggestion offers an improved disambiguation treatment.

2 Preliminaries

DOP uses a decomposed corpus as a grammar. When creating such a grammar, two decomposition operations, namely *Root* and *Frontier* are applied to all trees used for training. The former takes any node of a tree T and turns it into the root of a new subtree, erasing all nodes of T but the selected one and the ones it dominates. The new subtree is, then, processed by *Frontier*, which selects a set of nodes other than its root and erases all subtrees these dominate. Take for example the tree in 1(a) below.

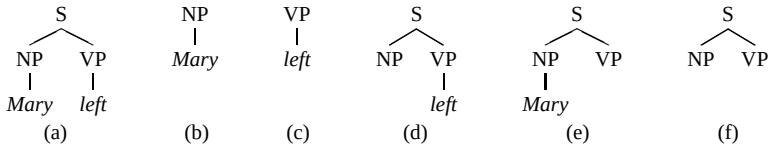

Figure 1: Example subtrees produced during training.

Applying *Root* to it will produce trees 1(a), (b) and (c). Applying *Frontier* to it will produce subtrees 1(a), (d), (e) and (f). *Frontier* does not have any effect on trees 1(b) and 1(c) since their non-root nodes (i.e. "Mary" and "left" respectively) do not dominate anything. Derivation, then, takes place by means of *leftmost substitution*. This composition operation substitutes the leftmost nonterminal leaf node L of some subtree t with another subtree rooted at L. For example, leftmost substitution of the subtrees 1(f) and 1(b) will yield a copy of the tree in 1(e).

Throughout the rest of this paper we will use the term *constituent* to refer to complete trees (complete in that all their leaf nodes are terminals). The fragments output by *Root*, for example, are constituents. The term *subtree* will be used to refer to some fragment whose leaf nodes might be terminal, nonterminal, or a combination of the two (as is the output of *Frontier*). The terms *corpus* and *treebank* will be used to refer to a collection of *constituents* and *subtrees*, respectively.

3 Bod's Probability Model

The fact that DOP, at least in its first instantiations, made use of syntactically labeled phrase structure trees to represent utterance analyses triggers high ambiguity in the syntactic structures produced. As a result, when parsing a new input string it is often the case that more than one analysis is generated. A probability model is, therefore, essential in order to disambiguate these analyses by providing an estimate for the most probable one. The model presented below is based on a simple frequency counter. Starting with the probability calculation of a certain derivation, two statistical assumptions have to be made with respect to the data to ensure its proper application:

1. *the elements in the treebank are stochastically independent*, and
2. *the treebank represents the total population (not just a sample) of subtrees*.

Let t be an X-rooted fragment in the Treebank. The probability of t being used at some stage of the derivation is given by the ratio of its frequency of occurrence over the frequency of occurrence of all X-rooted fragments in the treebank.

$$P(t) = \frac{|t|}{\sum_{i=1}^{n} f_{r(f)=X}}, \tag{1}$$

where $|t|$ is the frequency of t and $f_{r(f)=X}$ is the frequency of all X-rooted subtrees in the Treebank.

Suppose, now, we have a certain derivation d defined as $d = t_1 \circ t_2 \circ ... \circ t_n$. So long as the fragments are stochastically independent, the probability $P(d)$ of the derivation is defined as the product of the probabilities of the individual fragments being used (eq (2)).

$$P(d_j) = P(t_{1j}) \times P(t_{2j}) \times ... \times P(t_{nj}) = \prod_{i=1}^{n} P(t_{ij}) \tag{2}$$

Equation (1) defines the way of calculating the probabilities of each fragment in the corpus, and (2) how these can, then, be used to calculate the probability of a particular derivation. In some stochastic models that do not allow for fragments of arbitrary size, and especially in cases where the subtree depth is limited to one, the probability of a parse tree is identified with the probability of a single derivation (Charniak (1997)). In DOP, however, assuming all derivations d_j of some parse tree R are mutually exclusive, the probability of R is simply the sum of the probabilities of its individual derivations.

$$P(R) = \sum_{j=1}^{m} d_j = \sum_{j=1}^{m} \prod_{i=1}^{n} P(t_{ij}) \tag{3}$$

4 Bonnema and Scha's Probability Model

The problem with the above presented probability model is that it is biased and inconsistent (Johnson (2002)). The probability mass assigned to the subtrees arising from large corpus trees is overwhelmingly large. A well known example from the literature is that of Bonnema and Scha (2003), who show that in a treebank consisting of 1000 balanced binary trees of category S, 999 out of which are of depth 5 and 1 of depth 6, 99.8% of the overall probability mass for S goes to the descendents of the tree of depth 6.

On the basis that trees in the corpus do not carry any information about the subtree-probability pairs that were used in their derivation an alternative probability model was suggested by Bonnema and Scha (2003). The new model is based on Laplace's principle of insufficient reason, which states that in the absence of information regarding a set of solutions all alternatives should be assigned equal probabilities. Taking the set of derivations of some corpus tree τ to denote the set of alternative solutions, all derivations of τ in the corpus are considered a priori equally likely.

In order to calculate the substitution probability of a particular subtree, Bonnema and Scha (2003) consider a uniform distribution over derivations of a single tree τ. Let t be a subtree (to be used as an initial subtree in deriving τ), and $\delta(\tau)$ the set of all possible derivations of τ. Then the substitution probability of t, which is given by the probability distribution $\phi(t, \tau)$, is defined as the number of derivations of τ that start with t divided by the total number of derivations of τ.

$$\phi(t, \tau) = \frac{|d \in \delta(\tau) : d = t \circ ...|}{|\delta(\tau)|} \tag{4}$$

Moreover, let $N(\tau)$ denote the number of non-root non-terminal nodes (i.e. internal and substitution nodes) in τ. Then $|\delta(\tau)| = 2^{N(\tau)}$ (i.e. the number of possible derivations of τ equals the cardinality of

the powerset of the set of non-root non-terminal nodes of τ). If τ has $N(\tau)$ substitution nodes, and t has $N(t)$, then after t is substituted $N(\tau) - N(t)$ substitution nodes remain available. Substituting in Eq (4), we get the following:

$$\phi(t, \tau) = \frac{2^{N(\tau)-N(t)}}{2^{N(\tau)}} = 2^{-N(t)} \tag{5}$$

Eq (5) states that the substitution probability of t depends on its complexity alone. Since a uniform distribution of the derivations of a single tree was assumed, it follows that:

$$\sum_{t \in \sigma(\tau)} 2^{N(t)} = 1$$

When generalising over every tree in the Treebank, the total probability mass (i.e. 1) of each root category in the *corpus* is divided among its members in such a way that each one receives a proportion analogous to its relative frequency of occurrence. The probability mass $p(c) = f(c)/f(r(c))$ of a corpus tree c is then divided among its descendents t_i (i.e. the subtrees it gives rise to) by assigning a weight of $2^{-N(t_i)}$ of $p(c)$ to each one. The probability of a subtree t_i of c, hence, becomes:

$$p(t_i) = 2^{-N(t_i)} p(c) = 2^{-N(t_i)} \frac{f(c)}{f(r(c))}$$

where $f(c)$ and $f(r(c))$ are the frequencies of c and all corpus trees whose root node label is the same as that of c in the *corpus*. It is possible, however, for t_i to be a subtree of some other corpus tree c' as well, so the probability of t_i becomes:

$$p(t_i) = 2^{-N(t_i)} \frac{f(t_i)}{f(r(c))} \tag{6}$$

where $f(t_i)$ is the frequency of t_i in the *treebank*. The probability of a derivation d_j hence becomes:

$$p(d_j) = \prod_{i=1}^{m} 2^{-N(t_i)} \frac{f(t_i)}{f(r(c))} \tag{7}$$

and that of a parse tree T,

$$p(T) = \sum_{j=1}^{n} \prod_{i=1}^{m_j} 2^{-N(t_{ij})} \frac{f(t_{ij})}{f(r(c))} \tag{8}$$

With regard to the previously mentioned example, the new probability model assigns the more sensible figure of over 99.9% of the total probability mass of category S to the subtrees of the 999 constituents of depth 5. Only 0.1% of the probability mass is left for the subtrees of the constituent of depth 6, indicating that the bias in favour of large corpus trees has disappeared in this model.

5 The problem

Both probability models described above make use of the notion of a sample space containing only trees of the same category. Assuming each root category identifies its own sample space, however, leads to the following fundamental question. How do we compare trees belonging to different categories? This, in its turn, leads to an even more fundamental question. Do we really need to compare trees belonging to different categories? A real life example addressing this issue would be the following: According to a predefined DOP grammar, are we more likely to parse a given word like *book* as a noun or as a verb? Or even, are we more likely to parse a constituent like *her leaving school* as a DP, or as a VP? In certain domains the need for cross categorical comparison is greater. In human disambiguation this seems to be strongly related to the amount of context available. Phrases such as film, song or article titles, that appear out of context seem to present a higher degree of ambiguity. The film title “Supersize me”, for

example could either be a VP or an AP. Assuming that the mere existence of these situations identifies the need for cross-categorical comparison, we will turn to finding an answer to the first question.

We will first illustrate how the disambiguation algorithms discussed in the previous two sections assign to corpus trees probabilities disanalogous to what is expected, based on their observed relative frequency and known facts about *independence* of events. The remainder of this discussion is based on the following observation. Relative frequencies of constituents are identified with respect to the corpus being the sample space of alternatives, so they sum up to 1. Their respective probabilities, on the other hand, are calculated on the basis of some root-identifiable part of the corpus being the sample space of alternatives so they sum up to n, where n is the number of distinct root categories in the corpus. It is clear, even at this early stage of the discussion, that since the numbers associated with subtrees do not sum up to one, they are not *true* probabilities.

The relative frequency of a subtree with respect to any root-related identifiable sample space is not always enough for defining its probability. While after initiating the derivation process the root category of the next tree to take part in the derivation is known in advance, this is not the case when it comes to selecting the first subtree to start the derivation. In this instance, we do not have any information regarding the potential root category of the tree to be selected. It seems plausible, therefore, to assume that the sample space of alternatives at this point is the entire treebank.

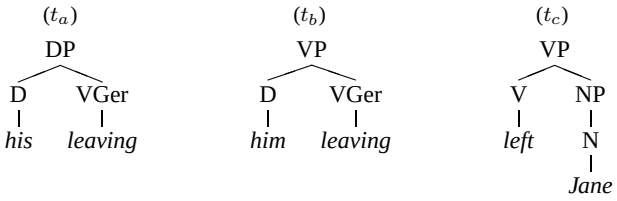

Figure 2: Initial trees

Consider, for example, the trees in Fig 2 giving rise to the treebank in Fig 3. For each subtree we have calculated its probability both according to the Bod (P_{Bod}) and the Bonnema and Scha ($P_{B\&S}$) model.

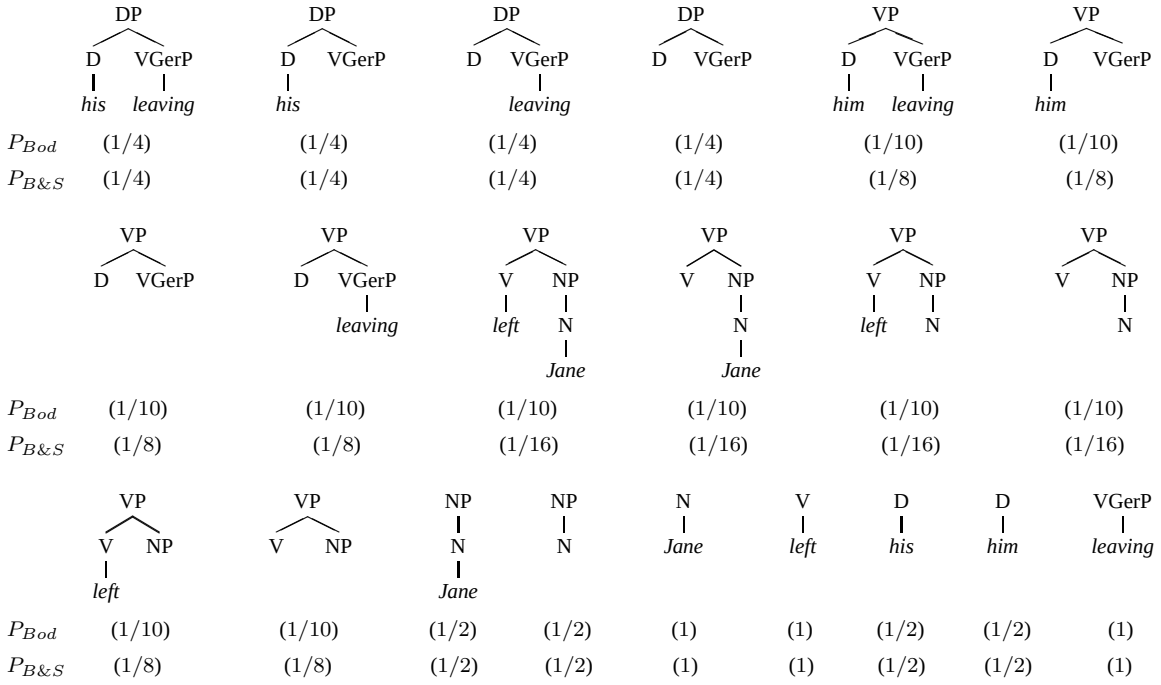

Figure 3: Treebank produced from the initial trees in Fig. 2

Even though the corpus does not meet the *independence constraint*, which states that the application of a rewrite rule or the selection of a subtree in this case is an independent event, two conclusions can be drawn about the probabilities of t_a, t_b and t_c by observing the data. First t_c does not exhibit any internal dependencies. In a PCFG analogy, this translates as: the fact that ($VP \rightarrow V\ NP$) is selected does not affect the probability of neither V being expanded as *left* nor NP being expanded as N. The probability assigned to tree t_c, therefore, is expected to be equal to its observed relative frequency. Trees t_a and t_b, on the other hand, do exhibit internal dependencies. D is a lot more likely to be expanded as *his* rather than *him* given that the rule ($DP \rightarrow D\ VGer$) has been selected. This positive tendency of ($DP \rightarrow D\ VGer$) and ($D \rightarrow$ *his*) occurring together, however, is equal to the positive tendency of ($VP \rightarrow D\ VGer$) and ($D \rightarrow$ *him*) occurring together. As a result, the probabilities assigned to t_a and t_b are expected to be equal given that their observed relative frequency is the same. Table 1, however, shows that neither model reflects the expected observations.

	(t_a)	(t_b)	(t_c)
Relative frequency	1/10	1/10	1/10
P_{Bod}	3/4	3/10	6/10
$P_{B\&S}$	3/4	3/8	4/8

Table 1: Relative frequencies and probabilities assigned by both models to t_a, t_b and t_c.

Clearly, the fact that VP is a more densely populated root category disrupts the way probabilities are assigned to constituents of different categories. This disruption can raise misclassification issues. Take, for example, the corpus depicted in Fig 4, assuming (t_2) occurs three times.

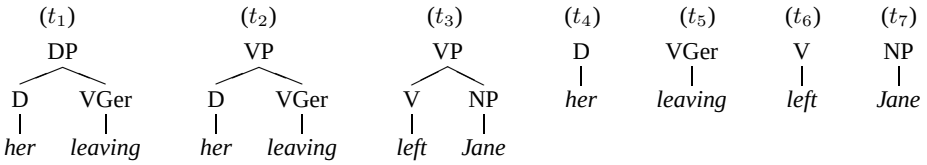

Figure 4: Corpus resulting from the initial trees t_1, t_2 and t_3.

For treebanks produced from this type of corpus the two models coincide in their predictions. They both assign exactly the same probabilities to all subtrees of the initial trees. Each of the four subtrees of t_1 and t_2 will have a probability of $1/4$ and $3/16$ respectively, while the probability of both (t_4) and (t_5) is 1. There are four ways of deriving the trees (t_1) and (t_2), which makes their corresponding probabilities 1 and $3/4$ respectively. As a result, both models wrongly predict analysis t_1 for the string *"her leaving"*, even though t_2 is three times more frequent in the training corpus.

This effect relates to the observation made earlier in this section that no single root identifiable sample space of alternatives is appropriate for initiating the derivation process. When selecting the first tree, the sample space includes all subtrees in the treebank. Once this tree has been selected, derivation is processed in some predefined manner (e.g. leftmost substitution in Bod (1995) or incrementally left to right in Neumann (2003)). At each step of the derivation a single substitution node is expanded. Its label, X, serves to identify a new sample space by selecting that part of the corpus containing only trees rooted at X.

The probabilities seen so far are conditional; hence they constitute a reasonable measure of comparison only if they are interpreted as such. $3/4$ is not, therefore, the probability of t_2. Rather, given that a tree starts with VP, it shows its likelihood of being this particular tree. Similarly, given that the tree to be derived starts with DP, its probability of being t_1 is 1 (the particular treebank cannot generate any DP-rooted trees other than this one). Each of the sample spaces identified by some category has a total probability mass of 1. As a consequence, comparison of tree probabilities is only meaningful

within the limits of each of these root identifiable sample spaces. What we have so far been calling probabilities are, in essence, just weights outside these limits.

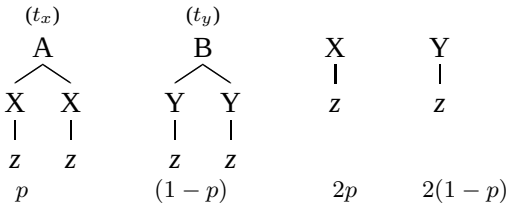

Figure 5: Corpus resulting from the initial trees t_x and t_y

In what follows we will demonstrate that the current model is not capable of generating trees with true probabilities. Suppose we have a sample of size n consisting of trees t_x and t_y above, with relative frequencies $w(t_x) = p$ and $w(t_y) = (1 - p)$, where p can take any value from 0 to 1. Given that the resulting corpus meets the *independence constraint*, the trees should be generated with probabilities $P_{cond}(t_x) = p/3$ and $P_{cond}(t_y) = (1 - p)/3$, which are the relative frequencies of t_x and t_y in the corpus depicted in Fig 5 (P_{cond} is used to refer to the Bonnema and Scha (2003) definition of fragment probabilities). By definition,

$$P_{cond}(t_x) = \sum_{j=1}^{4} \prod_{i=1}^{n} 2^{-N(t_{ij})} \frac{f(t_{ij})}{f(r(t_{ij}))} = \sum_{j=1}^{4} \prod_{i=1}^{n} \frac{1}{4} \frac{f(t_{ij})}{f(r(t_{ij}))} = \frac{1}{4} \sum_{j=1}^{4} \prod_{i=1}^{n} \frac{f(t_{ij})}{f(r(t_{ij}))}$$

In this example, however, $f(t_{ij}) = f(r(t_{ij})) = pn$ because there is no other constituent in the corpus with the same root as t_x. As a result, regardless of the frequencies of t_x and t_y in the corpus, t_x is always going to be assigned a probability of 1 ($P_{cond}(t_x) = 1$). This is due to the fact that the probability mass assigned to each root category is 1 and t_x has no other tree in its own sample space to share this probability with. One might wonder about the probability of t_y, which seems to be 0. This, however, reflects the probability of t_y with respect to the sample space identified by category A (i.e. $P(t_y|A)$), and since t_y is not of category A, its probability is naturally zero. If we change our working sample space from the one identified by A to the one identified by B, the probability of t_y becomes 1 ($P_{cond}(t_y) = P(t_y|B) = 1$) and that of t_x 0.

6 Putting all subtrees under the same category

The question of how to compare trees cross-categorically remains. The above discussion suggests that DOP is not equipped to handle such a task. One way of attacking the problem would be to incorporate a pseudo start-symbol, Z, in the design of the grammar. The reason Z is a pseudo start symbol is that it is not a member of the set of nonterminal symbols identifying the grammar. It is an extra symbol employed to simply mark constituent completeness and it does not bare any category related meaning at all. If the pseudo start symbol is formally incorporated in the grammar, parsing will produce, for each input string, pairs of almost identical parse trees, where one element will include the start symbol and one will not. A parse tree will then be considered *complete* iff it is anchored at exactly all the items in the input and it is rooted at the pseudo start symbol Z (for ease of reference we will refer to Z simply as a start symbol henceforth).

Let us reconsider the previous example assuming the annotation of initial trees is enriched with some start symbol, Z. The new corpus of trees t_x and t_y is presented in Fig 6 below.

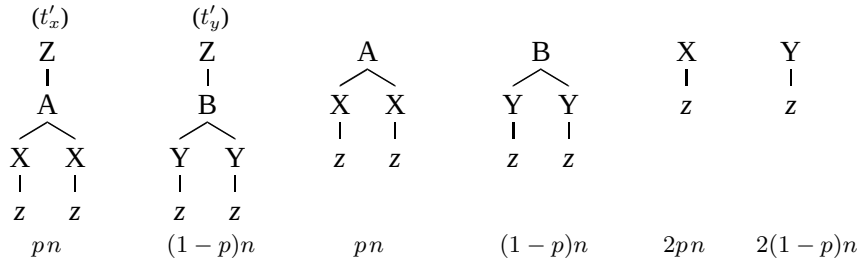

Figure 6: Corpus of initial trees incorporating a start symbol and their associated frequencies.

Again, given that the corpus meets the independence constraint, we expect t'_x and t'_y, the new counterparts of t_x and t_y, to be generated with weights equal to their respective observed relative frequencies ($p/3$ and $(1-p)/3$ as mentioned above).

$$P_{cond}(t'_x) = \sum_{j=1}^{8}\prod_{i=1}^{n} 2^{-N(t_{ij})}\frac{f(t_{ij})}{f(r(t_{ij}))} = \sum_{j=1}^{4}\prod_{i=1}^{n} 2^{-3}\frac{p\,n}{pn+(1-p)n} + \sum_{j=5}^{8}\prod_{i=1}^{n} 2^{-1}\frac{p\,n}{pn+(1-p)n}2^{-2}1 = p$$

Similarly, $P_{cond}(t'_y) = (1-p)$. On the one hand, adopting the start symbol approach seems to provide a very simple and straightforward way of comparing trees cross-categorically, by creating yet another sample space that accommodates complete constituents of different categories. Note, however, that in the example just examined it is possible to compare the different analyses of the string "z z", but not those of "z". This is due to the fact the analyses of the latter are not viewed as complete, so their probabilities are still identified with respect to different sample spaces. This, in its turn is caused by the start symbol being introduced at the level of initial trees. Moreover, the probabilities assigned to t'_x and t_y are still not equal to their observed relative frequencies.

In a framework such as DOP, however, where the treebank serves as a grammar, special attention should be paid to the way the start symbol Z is incorporated. If this approach is adopted, Z should be formally integrated in the grammar creation process, in order to reflect the fact that all initial trees and all their constituents are *complete* fragments. Since the output of *Root* when decomposing initial trees explicitly spells out the complete constituents seen in the training sample, it is there that Z should be introduced. Under this view, a corpus would contain all constituents produced by *Root* and a copy of each of these rooted at the start symbol. The corpus in Fig 6 now becomes as in Fig 7.

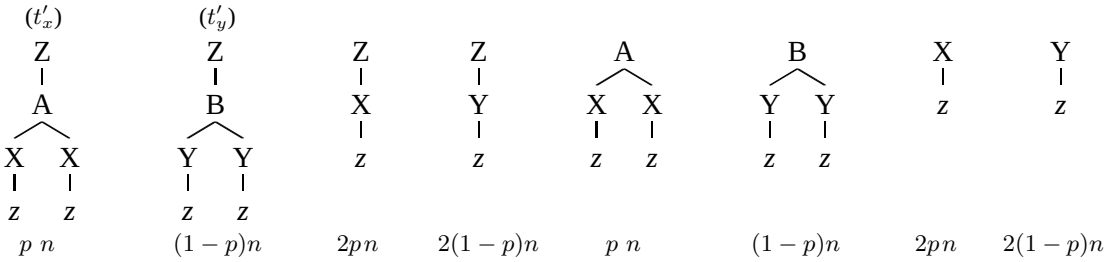

Figure 7: Corpus of constituents incorporating a start symbol and their associated frequencies.

Recalculating the probabilities shows that:

$$P_{cond}(t'_x) = \sum_{j=1}^{8}\prod_{i=1}^{n} 2^{-N(t_{ij})}\frac{f(t_{ij})}{f(r(t_{ij}))} = \sum_{j=1}^{4}\prod_{i=1}^{n} 2^{-3}\frac{p\,n}{pn+(1-p)n+2pn+2(1-p)n} +$$
$$+\sum_{j=5}^{8}\prod_{i=1}^{n} 2^{-1}\frac{p\,n}{pn+(1-p)n+2pn+2(1-p)n}2^{-2}1 = \frac{4}{8}\frac{p\,n}{3n} + \frac{4}{8}\frac{p\,n}{3n} = \frac{p}{3}$$

Similarly, $P_{cond}(t'_y) = (1-p)/3$. This formal incorporation of the start symbol as part of the grammar creation process thus provides an adequate way of addressing the issues discussed so far. It enables cross-categorical comparison of constituents that belong to different categories, and it assigns to trees probabilities that agree with the assumptions made based on their observed relative frequencies. The only problem is that the grammar more than doubles in size. While a set of m initial trees produce a treebank of size n without a start symbol, the same set of initial trees produces a treebank of size greater than $2n$ with a start symbol. Among other things, this otherwise unjustifiable increase in the number of subtrees would have a significant negative impact on the computational costs of processing new input as well as the ambiguity of the resulting grammar.

7 Redefining the probability of a derivation

An alternative to the approach discussed above comes from considering the issue of known vs. unknown information at each derivation step. When initiating the derivation process nothing is known about the potential root category of the first subtree to be selected. In other words, the first subtree f_1 is selected with respect to the sample space being identified by the treebank. Each subsequent fragment f_i taking part in the derivation, on the other hand, is selected with respect to the sample space identified by all fragments having the same root as f_i (i.e. the sample space is some root-identifiable part of the treebank). It seems, therefore, that two probabilities should be associated with each subtree to reflect what is known before its selection; an *absolute* probability which will be used for derivation initial selection and a *conditional* one which will be used if the subtree is selected later on in the derivation process. In what follows we will investigate how the probability of a derivation, and consequently that of a parse tree, will be affected by this argument. From this point onwards we will use the term *conditional probability* (P_{cond}) to refer to the probability of a subtree as defined in Bonnema and Scha (2003) and the term *absolute probability* (P_{abs}) to refer to the probability of a subtree with respect to the sample space identified by the corpus. We will also use P_{cond} to refer to the probability of a derivation or parse tree as defined in the literature so far, and P_{abs} to refer to the new definition of the probability of a derivation or parse tree that makes use of the *absolute* probability of the initial subtree.

Let us now calculate the *absolute probability* $P_{abs}(T)$ of a parse tree T of category X. The training corpus consists of n trees, n_x out of which are rooted at X. According to Bonnema and Scha (2003):

$$P_{cond}(t_{ij}) = P(t_{ij}|X) = 2^{-N(t_{ij})}\frac{f(t_{ij})}{n_x} \tag{9}$$

Moreover, according to the definition of conditional probability:

$$P(t_{ij}|X) = \frac{P(t_{ij} \cup X)}{P(X)} = \frac{P_{abs}(t_{ij})P(X|t_{ij})}{P(X)} = \frac{P_{abs}(t_{ij})}{P(X)} \tag{10}$$

Substituting 9 into 10 we get:

$$P_{abs}(t_{ij}) = 2^{-N(t_{ij})}\frac{f(t_{ij})}{n_x}P(X) = 2^{-N(t_{ij})}\frac{f(t_{ij})}{n_x}\frac{n_x}{n} = 2^{-N(t_{ij})}\frac{f(t_{ij})}{n} \tag{11}$$

Note that the *absolute probability* of a subtree depends on its size, its frequency and the size of the corpus. The subtrees taking part in each derivation d_j are t_{ij}, with t_{1j} being the initial subtree of d_j. The probability of a derivation $d_j = t_{1j} \circ ... \circ t_{mj}$ is now defined as the product of the *absolute probability* of the initial subtree and the *conditional probabilities* of all subsequent subtrees.

$$
\begin{aligned}
P_{abs}(d_j) &= P_{abs}(t_{1j}) \prod_{i=2}^{m} P_{cond}(t_{ij}) \\
&= 2^{-N(t_{1j})} \frac{f(t_{1j})}{n} \prod_{i=2}^{m} 2^{-N(t_{ij})} \frac{f(t_{ij})}{f(r(t_{ij}))} \\
&= \prod_{i=2}^{m} 2^{-N(t_{1j})} \frac{f(t_{1j})}{n} \frac{f(r(t_{1j}))}{f(r(t_{1j}))} 2^{-N(t_{ij})} \frac{f(t_{ij})}{f(r(t_{ij}))} \\
&= \frac{f(r(t_{1j}))}{n} \prod_{i=1}^{m} 2^{-N(t_{ij})} \frac{f(t_{ij})}{f(r(t_{ij}))} \\
&= P(r(t_{1j})) P_{cond}(d_j) \;=\; P(X) P_{cond}(d_j)
\end{aligned}
$$

The above shows how in order to calculate the *absolute probability* of a derivation we do not really need to calculate the *absolute probability* of the initial tree taking part in the derivation process. $P_{abs}(d_j)$ only depends on the conditional probability (P_{cond}) of d_j and the probability of its root node category. Consequently, the probability of T becomes:

$$
P_{abs}(T) = \sum_{j=1}^{k} P(d_j) = \sum_{j=1}^{k} P(X) P_{cond}(d_j) = P(X) \sum_{j=1}^{k} P_{cond}(d_j) = P(X) P_{cond}(T) \qquad (12)
$$

Similarly, the *absolute probability* of T depends on the conditional probability of T and the probability of its root node category.

In order to see the effect of this modification, we will reconsider some of the examples presented in this section. With respect to the first example at the beginning of the section (Fig. 2), the new probabilities become $P_{abs}(t_a) = P(\text{DP}) P_{cond}(t_a) = (1/10)(3/4) = 3/40$ and $P_{abs}(t_b) = P(\text{VP})\ P_{cond}(t_b) = (2/10)(3/8) = 3/40$, which confirms our observation that the two probabilities should be equal. Moreover, the probability of t_c, $P_{abs}(t_c) = P(\text{VP})\ P_{cond}(t_c) = (2/10)(4/8) = 1/10$, is equal to its observed relative frequency, as anticipated, since t_c presents no internal dependencies. Similarly, in the second example (Fig. 4) $P_{abs}(t_1) = P(\text{DP})\ P_{cond}(t_1) = (1/15)1 = 1/15$ and $P_{abs}(t_2) = P(\text{VP})\ P_{cond}(t_2) = (4/15)(3/4) = 3/15$, reflecting the fact that t_2 is three times more frequent in the training corpus than t_1. Turning now to the last example (Fig 5) the new probabilities are as follows:

$$
P_{abs}(t_x) = P(A) P_{cond}(t_x) = \frac{pn}{pn + (1-p)n + 2pn + 2(1-p)n} 1 = \frac{p}{3}
$$

$$
P_{abs}(t_y) = P(B) P_{cond}(t_y) = \frac{(1-p)n}{pn + (1-p)n + 2pn + 2(1-p)n} 1 = \frac{1-p}{3}
$$

The new probabilities accurately reflect what was expected (they are equal to the relative frequencies of the corresponding trees) for any value of p.

8 Concluding remarks

In this paper we have shown that even though the probability model proposed by Bonnema and Scha (2003) is consistent in its disambiguation predictions within the limits of a sample space identified by a certain root category, it cannot provide the necessary ground for cross-categorical comparison of trees by itself. The same problem is also present in the model described by Bod (1995). We briefly discussed

how formally incorporating a pseudo start symbol in the design of the corpus can offer the necessary grounds for disambiguation across categories. We moved on to arguing that, even though this approach solves the problem satisfactorily, it raises the size of the grammar to more than twice its original size, causing the computational costs of processing new input as well as the ambiguity inherent to the system to increase significantly. Moreover, there are no linguistic foundations backing up its existence.

As an alternative we proposed redefining the probability of a derivation in DOP, in order to reflect the fact that the root node label of the first constituent to be selected is unknown when initiating the derivation process. This approach translates the category specific conditional probability of a given derivation into category independent by assigning a weighted amount of the overall probability mass of the corpus (and hence the treebank) to each root node category. This approach is both linguistically and statistically more sound. In addition, it does not suffer from any additional processing costs while at the same time it does not affect the ambiguity of the resulting treebank. Even though this suggestion has not yet been evaluated in practice, it provides an adequate theoretical account for the issue described.

Acknowledgements

I would like to thank Dr Doug Arnold and Dr Aline Villavicencio for all their help and valuable comments. I would also like to thank the anonymous reviewers for their useful suggestions on how to improve this paper. Any flaws remain my own responsibility.

References

Bod, R. (1992) A Computational Model of Language Performance: Data Oriented Parsing, *Proceedings COLING'92* Nantes, France.

Bod, R. (1995) *Enriching Linguistics with Statistics: Performance Models of Natural Language*, PhD Thesis, Universiteit van Amsterdam, The Nehterlands.

Bonnema, R. and Buying, P. and Scha, R. (1999) A new probability model for Data Oriented Parsing, in P. Dekker and G. Kerdiles (eds.) *Proceedings of the 12th Amsterdam Colloquium*, Amsterdam, The Netherlands, 85-90.

Bonnema, R. and Scha, R. (2003) Reconsidering the Probability Model for DOP, in R. Bod, R. Scha and K. Sima'an (eds.) *Data-Oriented Parsing* (Stanford: CSLI), 25-42.

Charniak, E. (1997) Statistical parsing with a context-free grammar and word statistics, *Proceedings of AAAI'97,* (Menlo Park: AAAI Press/MIT Press), 598-603.

Johnson, M. (2002) The DOP Estimation Method is Biased and Iinconsistent, *Computational Linguistics 28,* (Menlo Park: AAAI Press/MIT Press), 71-76.

Neumann, G. (2003) A Data-Driven Approach to Head-Driven Phrase Structure Grammar, in R. Bod, R. Scha and K. Sima'an (eds.) *Data-Oriented Parsing* (Stanford: CSLI), 233-251.

Animacy classification based on morphosyntactic corpus frequencies: some experiments with Norwegian nouns

Lilja Øvrelid
NLP-unit, Dept. of Swedish
Göteborg University
lilja.ovrelid@svenska.gu.se

Abstract

This paper presents results from experiments in automatic classification of animacy for a set of Norwegian nouns using decision-tree classifiers. The method makes use of seven linguistically motivated syntactic and morphological features of these nouns extracted from an automatically annotated corpus of Norwegian. It classifies unseen nouns and achieves an accuracy of 90% under 10-fold cross-validation, as well as single hold-out training and testing.

1 Introduction

Animacy is a an inherent property of the referents of nouns which has been claimed to figure as an influencing factor in a range of different grammatical phenomena in various languages. It is also correlated with central linguistic concepts such as agentivity and discourse salience. Knowledge about the animacy of a noun is therefore relevant for several different kinds of NLP applications ranging from coreference resolution to parsing and generation. This article presents experiments on automatic classification of the animacy of unseen nouns, inspired by a method for verb classification, as presented in Merlo and Stevenson (2001). The experiments make use of statistical distributions of a set of linguistically motivated morphosyntactic cues for animacy, as gathered from an automatically annotated corpus of Norwegian.

In recent years a range of linguistic studies have examined the influence of argument animacy in grammatical phenomena such as differential object marking (Aissen, 2003), the passive construction (Dingare, 2001), the dative alternation (Bresnan et al., 2005) etc. A variety of languages are sensitive to the dimension of animacy in the expression and interpretation of core syntactic arguments (Lee, 2002; Øvrelid, 2004). Within the typological field of linguistics, the notion of *markedness* has been of great importance, underlying much of the cross-linguistic comparative work performed there. This notion is based on "asymmetrical or unequal grammatical properties of otherwise equal linguistic elements"(Croft, 2003), and is linked to, among others, the relative frequency of a given structure. An unmarked structure will typically be more frequent than its marked counterpart and relatedly, figure in a greater number of linguistic contexts (Croft, 2003). So-called prominence hierarchies figure frequently in typological descriptions. These hierarchies express the relative prominence of a structure, and incorporate the relativity of markedness into the theory. The notion of prominence has been linked to several properties such as most likely as topic, agent, most available referent etc. Among the hierarchies established in typological literature are those of syntactic functions, animacy and thematic role (Croft, 2003; Aissen, 2003):

Syntactic function: Subject > Object

Animacy: Human > Animate > Inanimate

Thematic Role: Agent > Patient

A key generalisation or tendency regarding these hierarchies is that features placed high on one hierarchy tend to attract other prominent or high-placed features; subjects, for instance, will tend to be animate and agentive, whereas objects prototypically are inanimate and themes/patients. Exceptions to this generalisation express a more *marked* structure, a property which has consequences, for instance, in the distributional properties of the structure in question.

Even though knowledge about the animacy of a noun clearly has some interesting implications, little work has been done within the field of lexical acquisition in order to automatically acquire such knowledge. Orăsan and Evans (2001) make use of hyponym-relations taken from the WordNet resource in order to classify animate referents. However, such a method is clearly restricted to languages for which large scale lexical resources, such as the WordNet, are available. Merlo and Stevenson (2001) present a method for verb classification which relies only on distributional statistics taken from corpora in order to train a decision tree classifier to distinguish between three groups of intransitive verbs. A key question in the following thus becomes whether a similar method may be applied to the task of animacy classification based on linguistically motivated cues extracted from a corpus.

2 Morphosyntactic features of animacy

The method for verb classification described in Merlo and Stevenson (2001) makes use of a training set consisting of relative frequency data for each verb in a certain class, which summarise its overall count for a certain feature. The task is thus a matter of classifying an unseen verb based on properties of all the instances of this verb (the lemma), rather than classifying individual instances by themselves.

What features, then, can be exploited as cues for the animacy of a noun? As mentioned above, animacy is highly correlated with a number of other linguistic concepts, such as agentivity, topicality and discourse salience. One would expect marked configurations along these dimensions, e.g. animate objects or agentive inanimates, to be less frequent in the data. However, these are complex notions to translate into extractable features from a corpus. In the following we will present some morphological and syntactic features which, in different ways, approximate the multi-faceted property of animacy. It is important, however, to stress that these features only provide *approximations* of animacy, which, hopefully, lead to observable distributional differences between nouns.

As mentioned earlier, a prototypical transitive relation involves an animate subject and an inanimate object. In fact, a corpus study of animacy distribution in simple transitive sentences[1] in Norwegian revealed that approximately 70% of the subjects of these types of sentences were animate, whereas as many as 90% of the objects were inanimate (Øvrelid, 2004). Although this corpus study involved all types of nominal arguments, i.e. pronouns and proper nouns as well, it still seems that the frequency with which a certain noun occurs as a subject or an object of a transitive verb might be an indicator of its animacy.

Agentivity is another related notion to that of animacy, animate beings are usually inherently sentient, capable of acting volitionally and causing an event to take place - all properties of the prototypical

[1] Simple transitive sentences are main sentences which include a simple transitive main verb, i.e. no auxiliaries or modals. Due to the fact that Norwegian is a V2-language which does not case mark nouns, and allows for both SVO and OVS word order, sentences like these are syntactically/functionally ambiguous (Øvrelid, 2004).

agent, according to Dowty (1991). However, if no additional information on argument structure for verbs is to be assumed, other ways of approximating the agentivity of a noun must be arrived at. One possibility is to use the passive construction, or rather the property of being expressed as the demoted agent in a passive construction. As is well known, transitive constructions tend to passivise better (hence more frequently) if the demoted subject bears a prominent thematic role, preferably agent, rather than a role less prominent on the thematic role hierarchy. A prediction to be tested is therefore whether the relative frequency with which a noun occurs in a passive by-phrase, is an indicator of its animacy.

Anaphoric reference is a phenomenon where the animacy of a referent is clearly expressed. The Norwegian personal pronouns distinguish their antecedents along the animacy dimension - animate *han/hun* 'he/she' vs. inanimate *den/det* 'it-MASC/NEUT'. This is one reason why information regarding the animacy of a noun can be helpful in the task of coreference resolution. However, in this context it might be interesting to make use of an approximation of anaphoric reference in determining the animacy of a noun.

Reflexive pronouns represent another form of anaphoric reference, and, may, in contrast to the personal pronouns locate their antecedent locally, i.e. within the same clause[2]. The third person Norwegian reflexive pronoun *seg* 'him/her/itself' does not, however, differentiate its antecedent along the animacy dimension. In the prototypical reflexive construction the subject and the reflexive object are coreferent and it describes an action directed at oneself. Although the reflexive pronoun in Norwegian does not distinguish for animacy, the agentive semantics[3] of the construction might favour an animate subject.

Finally, when it comes to morphological properties, animate nouns are not marked specifically as such in Norwegian. Common nouns are marked for number and definiteness, as well as having an inherent gender (masculine, feminine or neuter). There is no extensive case system for common nouns and the only distinction that is explicitly marked on the noun is the genitive case by addition of -s. The genitive construction typically describes possession, a relation which often involves an animate possessor. However, this is certainly not always the case, semantic relationships such as a whole-part relation as in *bilens hjul* 'the car's wheel' or a quantificational meaning as in *en times arbeide* 'an hour's work' etc. also commonly occur. An alternative construction to the s-genitive in Norwegian is constructed by inserting the possessive pronoun *sin* between the possessor and the possessed, as in *mannen sin bil* 'the man's car'. The *sin*-genitive is to be preferred when the relation is one of possession (Faarlund et al., 1997), hence often involving an animate possessor. Generally then, the frequency with which a noun occurs as a modifier might provide an indicator of the animacy of that noun.

3 Feature extraction

In order to train a classifier to distinguish between animate and inanimate nouns, training data consisting of distributional statistics have to be extracted from a corpus. Appropriate approximations of the linguistically motivated features described above also have to be constructed. For this end, a 15

[2]Norwegian has two types of reflexive constructions - a simple reflexive *seg* and a complex reflexive *seg selv*. The difference between these two have traditionally been viewed as based on locality - the complex reflexive is bound locally, whereas the simple one is bound non-locally. However, as Lødrup (1999) shows, this represents an idealization which does not hold up against real data. In particular, the simple reflexive pronoun is far more versatile than previously assumed and may very well be bound locally. We will therefore include both the simple and complex reflexives in our study.

[3]Reflexives in Norwegian do not necessarily express an agentive event, and may be employed, for instance in medial constructions. However, with regards to productivity one would assume the agentive relfexives to be predominant, hence implying an animate subject.

million word version of the Oslo Corpus, a corpus of Norwegian texts of approximately 18.5 million words, was employed[4]. The corpus is morphosyntactically annotated and assigns an underspecified dependency-style analysis to each sentence[5].

As training data for the classifier, a set of forty nouns were chosen - twenty animate and twenty inanimate nouns, examplified in (1a) and (1b) respectively:

(1) (a) *barn* 'child', *direktør* 'director', *far* 'father', *flyktning* 'refugee', *forfatter* 'author', *gutt* 'boy', *leder* 'leader', *lege* 'doctor'

(b) *aksje* 'stock', *artikkel* 'article', *bil* 'car', *bok* 'book', *brev* 'letter', *dag* 'day', *eiendom* 'property', *fly* 'airplane'

The corpus study of Norwegian simple transitives mentioned earlier, showed that nouns expressing animate beings aside from humans (e.g. animals) are very infrequent (0.0025%) in the corpus (Øvrelid, 2004), and these were therefore not focused on in the following. Also, as some of the features employed were assumed to be quite rare, e.g. anaphoric pronominal reference or passive by-phrases, a cut-off point with regards to frequency was maintained throughout the study; all nouns had at least one thousand occurrences in the corpus.

3.1 Feature approximation

For each noun, relative frequencies for the different morphosyntactic features described above were computed from the corpus.

Subjects and objects For transitive subjects, we extracted the number of instances where the noun in question was unambiguously tagged as subject and followed by a finite verb and an unambiguously tagged object[6]. The frequency of direct objects for a given noun was approximated to the number of instances where the noun in question was unambiguously tagged as object. We here assume that an unambiguously tagged object implies an unambiguously tagged subject. However, by not explicitly demanding that the object is preceded by a subject, we also capture objects with a "missing" subject, such as relative clauses and infinitival clauses.

Passive As we remember, another context where animate nouns might be predominant is in the by-phrase expressing the demoted agent of a passive verb. Norwegian has two ways of expressing the passive, a morphological passive (verb + *s*) and a periphrastic passive (*bli* + past participle). The counts for passive by-phrases allow for both types of passives to precede the by-phrase containing the noun in question.

Anaphoric reference With regards to the property of anaphoric reference by personal pronouns, the extraction was bound to be a bit trickier. The anaphoric personal pronoun is never in the same clause as the antecedent, and often not even in the same sentence. Coreference resolution is a complex problem, and certainly not one that we shall attempt to solve in the present context. However, we might attempt

[4]The corpus is freely available for research purposes, see http://www.hf.uio.no/tekstlab for more information.

[5]The actual framework is that of Constraint Grammar (Karlsson et al., 1995), and the analysis is underspecified as the nodes are labelled only with their function, e.g. subject or prepositional object, and not its immediate head or dependent(s).

[6]The tagger works in an eliminative fashion, so tokens may bear two or more tags when they have not been fully disambiguated.

to come up with a metric that approximates the coreference relation in a manner adequate for our purposes, that is, which captures the different coreference relation for animate as opposed to inanimate nouns. To this end, we make use of the common assumption that a personal pronoun usually refers to a discourse salient element which is fairly recent in the discourse. Now, if a sentence only contains one core argument (i.e. an intransitive subject) and it is followed by a sentence initiated by a personal pronoun, it seems reasonable to assume that that these are coreferent (Hale and Charniak, 1998). (2) below shows an authentic example from the results for the noun *mann* 'man' taken from the Oslo Corpus:

(2) **Mannen**$_i$ ble pågrepet etter tre kvarters dramatisk biljakt. **Han**$_i$ var beruset ...
The man$_i$ was apprehended after a three-quarter long car chase. He$_i$ was intoxicated ....

For each of the nouns then, we count the number of times it occurs as a subject with no subsequent object and an immediately following sentence initiated by (i) the animate personal pronouns *han* 'he', *hun* 'she' or *de* 'they', and (ii) the inanimate personal pronouns *den* 'it-MASC' or *det* 'it-NEUT'. Now, the 3. person plural *de* 'they' is not strictly an indicator of animacy as it may refer to both animate and inanimate referents, as in English. However, Merlo and Stevenson (2001) claim that, in English, this plural pronoun usually refers to animate entities and in a selection of 100 occurrences of this pronoun, they found that 76% of these had an animate antecedent[7]. We therefore make the same assumption for Norwegian, although this is a possible source for mistakes in the counts, we assume that the general distribution of instances will still differentiate with regards to animacy. Another possible source for mistakes in the relative frequencies lies in the fact that we cannot assume to have knowledge regarding the natural gender of our training nouns. As this often does not coincide with the grammatical gender of a noun in Norwegian, we must therefore count all occurrences of the personal pronouns following a noun without controlling for agreement with respect to natural gender.

For the inanimate pronouns, the neuter form *det* 'it-NEUT' is problematic as this is also the expletive subject form. This pronoun therefore often initiates a sentence, but has a clearly non-referential function. However, as the distinction between expletive and pronominal subjects is not annotated for in the corpus, we will count all occurrences of this pronoun when it initiates a subsequent sentence. Another possibility would have been to exclude all occurrences of *det* 'it' from the counts, with the consequence that this test would be inapplicable for the set of neuter nouns in our training set (8 nouns).

Reflexive The feature of reflexive coreference is easier to approximate, as this coreference takes place within the same clause. For each noun, the number of occurrences as a subject followed by a verb and the 3.person reflexive pronoun *seg* 'him-/her-/itself' are counted and its relative frequency recorded.

Genitive -s This feature simply contains relative frequencies of the occurrence of each noun with genitive case marking, i.e. the suffix *-s*. As mentioned earlier, the Norwegian *sin*-genitive is usually preferred with animate possessors and might provide a useful feature of animacy. Unfortunately, however, this construction is far too rare and yielded zero occurrences for a large number of the nouns (both animate and inanimate), hence was abandoned.[8]

[7] Merlo and Stevenson (2001) make use of personal pronouns as indicators of argument structure for a verb. If it often occurs with an animate pronominal subject, they assume that the verb distributes agentive role to its subject.

[8] The *sin*-genitive is generally a property of spoken rather than written Norwegian, although one can find examples in more informal writing (Faarlund et al., 1997).

	SUBJ		OBJ		GEN		PASS		ANAANIM		ANAINAN		REFL	
Class	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A	0.14	0.05	0.11	0.03	0.04	0.02	0.006	0.005	0.009	0.006	0.003	0.003	0.005	0.0008
I	0.07	0.03	0.23	0.10	0.02	0.03	0.002	0.002	0.003	0.002	0.006	0.003	0.001	0.0008

Table 1: Mean relative frequencies and standard deviation for each class (A(nimate) vs. I(nanimate)) from feature extraction (SUBJ=Transitive Subject, OBJ=Object, GEN=Genitive *-s*, PASS=Passive *by*-phrase, ANAANIM=Anaphoric reference by animate pronoun, ANAINAN=Anaphoric reference by inanimate pronoun, REFL=Anaphoric reference by reflexive pronoun).

3.2 Results

The mean relative frequencies for each class - animate and inanimate - are presented in table 1. The standard deviation for each class and feature is provided alongside the mean. The total data points for each feature following the data collection are as follows: SUBJ: 16813, OBJ: 24128, GEN: 7830, PASS: 577, ANAANIM: 989, ANAINAN: 944, REFL: 558. As we can see, quite a few of the features express morphosyntactic cues that are rather rare. This is in particular true for the passive feature and the anaphoric features ANAANIM, ANAINAN and REFL. This is perhaps not so surprising, however, the question is whether these features express the relevant distinction although they are sparse. When examining the features in table 1. this certainly seems to be the case; the difference between the mean feature values for the two classes range from double to five times the lowest class value.

Another point is that the values for the features that one would expect to be quite frequent, e.g. SUBJ and OBJ only range from about 3% to 14% of all occurrences. The reason for this is that the search patterns designed to extract the counts require the subjects and objects in question to be *unambiguously tagged*. However, all subjects and objects of the simple transitive sentences mentioned earlier are tagged as being both subjects and objects on account of their functional ambiguity. This means that the transitive subjects and objects that *are* counted are only those that occur in a syntactic environment which clearly disambiguates them functionally[9].

3.3 Other features

The features and the mean values presented in table 1. are the features that were actually employed in the experiments. However, several other features were also extracted, which did not exhibit the required distinction. The indirect object of the ditransitive double object construction expresses the thematic role of recipient and is known for displaying an "animacy effect" (Bresnan et al., 2005). An approximation was therefore attempted for the feature of indirect objects in ditransitive constructions. This however, turned out to yield a result that was contrary to the expected results. The mean result for the animate class was 0.007%, whereas the inanimate class had the higher count of 0.008%. However, a quick look at some of the extracted sentences shows that the tagger's automatic analysis of indirect objects contains a lot of errors. This feature was therefore abandoned and not included in the classification experiments.

Due to the mentioned correlation between animacy and discourse salience or topicality, the morpho-

[9]In practice this includes transitive complex VPs (due to the V2-property of Norwegian), i.e. VPs containing auxiliary or modal verbs, sentences where something other than the subject or object occupies sentence initial position, or subjects or objects occupying subordinate clauses of different types.

logical definiteness of the animate vs. inanimate nouns was also recorded. One might assume that a topical element is also definite. However, this feature only yielded a mean 1% difference between the categories, hence was also abandoned. One possible reason for this is that morphological and semantic definiteness do not necessarily overlap, hence the crude measure of morphological definiteness might not be able to fully capture the semantic definiteness of a given noun.

4 Experiments

The experimental methodology chosen for the classification experiments for animacy is pretty much identical to the one described in Merlo and Stevenson (2001) for verb classification. The same software package for decision tree learning, C5.0 (Quinlan, 1993), has also been employed[10].

A decision tree is a classification model which relates a set of predefined classes with properties of the instances to be classified. In this case we wish to classify Norwegian common nouns along the binary dimension of animacy, i.e. animate vs. inanimate. The properties in question are the morphosyntactic features on which we have gathered data. Classification using a decision tree proceeds by means of a set of weighted, disjunctive tests which at each step (node) in the process assigns an appropriate test to an input, and which proceeds along one of its branches, representing possible outcomes of the test.

4.1 Training and testing methodology

Based on the data collected on seven different features for our 40 nouns, a set of feature vectors may be constructed for each noun. They contain the relative frequencies for each feature along with the name of the noun and its class (animate or inanimate). Note that the vectors do not contain the mean values presented in table 1. above, but rather the individual relative frequencies for each noun.

Merlo and Stevenson (2001) experiment with two different methodologies for training and testing the decision tree classifier(s) - 10-fold cross-validation and single hold-out. They have in common that the reported results from both are on unseen test data, i.e. data that are not part of the training set, however they are also different in the sense that their results contribute slightly different information (Merlo and Stevenson, 2001). 10-fold cross-validation has the advantage that it reports an average accuracy result for the entire data set, whereas single hold-out provides more specific results regarding which classes and nouns are misclassified, thus forming the base for further analysis. For our experiments in animacy classification both methods for training and testing were employed. As the task is a binary classification task, we assume a baseline accuracy of at best 50%.

4.2 Results

4.2.1 10-fold cross-validation

As the 10-fold cross validation method reports accuracy measures averaged over all runs, it facilitates the testing of different features and their individual contribution to the classification task.

Table 2. shows the performance of each individual feature in the classification of animacy. As we can see, the features perform quite well, ranging from mere baseline performance (ANAINAN) to a 65% improvement of the baseline (REFL).

[10]The C5.0 software package may be downloaded from http://www.rulequest.com/.

Feature	% Accuracy
SUBJ	77.5
OBJ	72.5
GEN	75.0
PASS	67.5
ANAANIM	70.0
ANAINAN	50.0
REFL	82.5

Table 2: Accuracy for the individual features using 10-fold cross validation

Features used	Feature Not Used	% Accuracy
1. SUBJ OBJ GEN PASS ANAANIM ANAINAN REFL		90.0
2. OBJ GEN PASS ANAANIM ANAINAN REFL	SUBJ	85.0
3. SUBJ GEN PASS ANAANIM ANAINAN REFL	OBJ	90.0
4. SUBJ OBJ PASS ANAANIM ANAINAN REFL	GEN	85.0
5. SUBJ OBJ GEN ANAANIM ANAINAN REFL	PASS	82.5
6. SUBJ OBJ GEN PASS ANAINAN REFL	ANAANIM	77.5
7. SUBJ OBJ GEN PASS ANAANIM REFL	ANAINAN	85.0
8. SUBJ OBJ GEN PASS ANAANIM ANAINAN	REFL	77.5

Table 3: Accuracy for all features and 'all minus one' using 10-fold cross validation

The first line of table 3. shows the performance using all the seven features where we achieve an accuracy of 90%, an 80% improvement of the baseline. The subsequent lines of table 3. show the accuracy results for classification using all features except one at a time. This provides an indication of the contribution of each feature to the classification task. The removal of the transitive object feature in line 3. does not affect the accuracy of the classifier at all and this feature is therefore redundant. Removal of the transitive subject feature on the other hand, causes a 5% deterioration of accuracy.

In general, the removal of a feature causes a 0% - 12.5% deterioration of results. We also see that the behaviour of the features in combination is not strictly predictable from their individual performance, as presented in table 2. For instance, the removal of the 'anaphoric reference with animate pronoun' feature (ANAANIM) has the most severe effect on the result, but is one of the poorest performing features on its own.

4.2.2 Single hold-out

As mentioned earlier, the single hold-out method has the advantage of providing results regarding the individual classes as well as individual nouns. Because it facilitates class-wise comparisons, a F score may also be computed, which relates true/false negatives and positives for each class[11]. In this case, the simple accuracy (number of correct classifications / all classifications) and the F score are identical when all the features are employed, as shown in line 1 in table 4. The number of misclassifications are symmetrical - two nouns are misclassified for each class (hence two were deemed false positives for the opposing class). As we see, then, the result for all the features combined is the same as for the 10-fold cross validation method - 90% accuracy. For the 'all minus one' feature sets the results are not completely identical to that of the 10-fold cross-validation method. The accuracy measures differ somewhat, showing that the learner is slightly sensitive to the exact makeup of the test sets. Also, the removal of the object feature here shows a 2.5% deterioration of results, caused by the misclassification of one extra noun, in contrast to the cross-validation results. It seems that the SUBJ and OBJ features are somewhat overlapping. A possible reason for this, is that the information contained in the subject feature actually implies a direct object, only nouns that were unambiguously tagged as subject and followed by an unambiguous object were counted. As we remember, the object feature, on the other hand, does not demand a realized subject.

The balanced F score provided for each class in table 4. provides us with a more detailed picture of

[11]We make use of a balanced F score: 2PR/P+R (P=precision: true positives / true positives + false positives, R=recall: true positives / true positives + false negatives) (Merlo and Stevenson, 2001)

Features Used	Not Used	% Acc	% F Anim	% F Inan
1. SUBJ OBJ GEN PASS ANAANIM ANAINAN REFL		90.0	90.0	90.0
2. OBJ GEN PASS ANAANIM ANAINAN REFL	SUBJ	85.0	84.2	85.7
3. SUBJ GEN PASS ANAANIM ANAINAN REFL	OBJ	87.5	87.8	87.2
4. SUBJ OBJ PASS ANAANIM ANAINAN REFL	GEN	85.0	85.0	85.0
5. SUBJ OBJ GEN ANAANIM ANAINAN REFL	PASS	82.5	83.7	81.1
6. SUBJ OBJ GEN PASS ANAINAN REFL	ANAANIM	82.5	82.0	83.0
7. SUBJ OBJ GEN PASS ANAANIM REFL	ANAINAN	85.0	84.2	85.7
8. SUBJ OBJ GEN PASS ANAANIM ANAINAN	REFL	72.5	73.2	71.8

Table 4: Accuracy and balanced F-score per class for all features and 'all minus one' using single hold-out method.

the effect of each feature on each class, as measured by the removal of this feature from the feature set. We are also informed of whether the effects of the features are as we predicted earlier. This seems largely to be the case; the removal of a feature which targets a specific class causes a lower F score for this class. For instance, the removal of SUBJ causes a lower F score for the animate class than the inanimate, indicating that a higher number of misclassifications related to the animate class took place.

In general, it seems fair to say that more features perform better. The nouns that are misclassified following the removal of a feature are seldom the same ones, hence underlining the need for all the features. An idiosyncratic behaviour of a noun in the light of one specific feature, is attributed less importance when the evidence from all the features is weighted in.

5 Discussion

The above experiments have shown that the classification of animacy for common nouns is achievable using distributional data from a syntactically annotated corpus. The results of the experiments are encouraging, and due to the fact that the features are linguistically motivated, hopefully also generalisable to a larger set of nouns. However, several questions remain open for future work.

We have chosen to classify along a binary dimension (animate vs. inanimate) with a relatively small set of nouns. Two related objections may be put forward at this point. Firstly, it might be argued that a binary dimension such as this is artificial and that there should be a finer subdivision of nouns. Zaenen et al. (2004) describe an encoding scheme for the manual encoding of animacy information in part of the English Switchboard corpus. They make a three-way distinction between human, other animates, and inanimates, where the 'other animates' category describe a rather heterogeneous group of entities: organisations, animals, intelligent machines and vehicles. However, what these seem to have in common is that they may all be construed linguistically as animate beings, even though they, in the real world, are not. Interestingly, the two misclassified inanimate nouns in our experiments, were *bil* 'car' and *fly* 'airplane', both vehicles. They exhibited a more agentive pattern which showed up in the transitive subject feature, the passive feature and the reflexive feature, in particular. However, they did not pattern completely with the animate nouns, they had a high object count and behaved like the inanimate nouns when it came to anaphoric pronouns. Secondly and related to the above, the choice of nouns in the experiment might be considered too limited. Had we chosen to include, for instance, nouns that have a metonymic use e.g. organisations, the classification into only two classes might have been less successful. However, we chose to start out with a binary classification in order to test

the viability of the method and its suitability for the classification task. Further experiments should probably enlarge the set of training nouns and also include an intermediate category, as proposed in Zaenen et al. (2004).

One might also ask whether the chosen features represent sufficient information to base classification on. As mentioned several times, the features only provide approximations of animacy by relying on related linguistic dimensions such as syntactic functions and thematic roles. Now, one of the misclassified animate nouns was *venn* 'friend', a clearly animate noun. However, according to our seven chosen features, this noun largely patterns with the inanimate nouns. When considering it, this probably also makes sense, as we are basing our classification of a real world property only on our linguistic depiction of it. A friend is probably more like a physical object in the sense that it is someone one likes/hates/loves or otherwise reacts *to*, rather than being an agent that acts upon its surroundings. Also, it is neutral with regards to natural gender, hence, probably less likely to be followed by a gender-specific pronoun. The features for anaphoricity therefore point more in the direction of inanimate nouns, as well.

In the long run, the acquisition of animacy by itself is not necessarily the only goal. By testing the use of acquired animacy information in various applications, such as parsing, generation or coreference resolution, the generalisations from linguistic studies regarding animacy effects in human language may be made use of, or even tested. A common problem in studies that rely on corpus data, however, is data sparseness. As mentioned earlier, the nouns experimented with in this study are all rather frequent (more than a thousand occurrences) in the corpus data. However, if one wishes to scale up the current approach, this problem will have to be dealt with. As the cut-off point of one thousand occurrences was rather randomly selected, an experiment was performed where the cut-off point was drastically reduced to approximately one hundred occurrences (+/- 20). Ten nouns of each class were attempted classified by i) the classifier trained earlier on the more frequent nouns by single hold-out ii) a new classifier trained and tested only on the twenty infrequent nouns by single hold-out. Both the experiments showed that most of the features are affected negatively by sparse data, a result which is not at all surprising, given that quite a few of the features are rather rare. Both the experiments yielded an accuracy of 65% when all seven features were employed. An error analysis showed that the majority of mistakes made by the two classifiers were misclassifications of animate nouns as inanimate. Due to the fact that the majority of our features (SUBJ, GEN, PASS, ANAANIM, REFL) require a higher relative frequency for animates than inanimates, it seems obvious that these are the nouns which will suffer most from sparse data in classification. However, both the classifiers improved their performance quite drastically when tested with only the transitive subject feature SUBJ - a more frequent feature which targets animate nouns. The old classifier applied to the twenty low frequency nouns here achieved an accuracy of 80%, whereas the classifier trained only on the new nouns achieved an accuracy of 85%. It thus seems that backing off to more frequent features when classifying lower frequency nouns might be a strategy worth investigating further. Experiments should also be performed in order to locate an appropriate cut-off point, as well as investigating further the interaction of our features.

In conclusion then, we have seen that the method for verb classification described in Merlo and Stevenson (2001) yield promising results for classification of animacy when applied to Norwegian common nouns using a set of seven linguistically motivated features of animacy. The theoretical predictions that the relative markedness of a construction along so-called prominence hierarchies would influence its frequency turned out to provide useful clues for an automatic classification task.

References

Judith Aissen. Differential Object Marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21:435–483, 2003.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. Predicting the dative alternation. To appear in Royal Netherlands Academy of Science Workshop on Foundations of Interpretation proceedings, 2005.

William Croft. *Typology and Universals*. Cambridge University Press, 2nd edition, 2003.

Shipra Dingare. The effect of feature hierarchies on frequencies of passivization in English. Master's thesis, Stanford University, August 2001.

David Dowty. Thematic proto-roles and argument selection. *Language*, 67(3):547–619, 1991.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. *Norsk Referansegrammatikk*. Universitetsforlaget, 1997.

John Hale and Eugene Charniak. Getting useful gender statistics from English text. Technical report, Comp. Sci. Dept. at Brown University, Providence, Rhode Island, 1998.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Atro Anttila, editors. *Constraint Grammar: A language-independent system for parsing unrestricted text*. Mouton de Gruyer, 1995.

Hanjung Lee. Prominence mismatch and markedness reduction in word order. *Natural Language and Linguistic Theory*, 2002.

Helge Lødrup. Inalienables in Norwegian and binding theory. *Linguistics*, 37(3):365–388, 1999.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. The MIT Press, 1999.

P. Merlo and S. Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):373–408, 2001.

Constantin Orăsan and Richard Evans. Learning to identify animate references. In *Proceedings of the Workshop on Computational Natural Language Learning*. ACL-2001, 2001.

Lilja Øvrelid. Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. In Fred Karlsson, editor, *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki, 2004.

J. Ross Quinlan. *C4.5: Programs for machine learning*. Series in Machine Learning. Morgan Kaufmann Publishers, 1993.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. Animacy encoding in English: why and how. In D. Byron and B. Webber, editors, *ACl Workshop on Discourse Annotation*, Barcelona, 2004.

Building on Syntactic Annotation: Labelling of Subordinate Clauses

Marina Santini
ITRI
University of Brighton
Marina.Santini@itri.brighton.ac.uk

Abstract:

In this paper, we explore the possibility of labelling semantic/adverbial clauses (i.e. *temporal* clauses, *purpose* clauses, *concession* clauses, etc.), complement/nominal clauses (i.e. *ing*-clauses, *that*-clauses, etc.), and complex noun phrases (NPs) using syntactic patterns built on the syntactic annotation returned by a parser. We suggest that the use of syntactic patterns can represent a useful alternative among other methods for syntax extraction proposed so far. The approach presented here is parser-dependent, but it can be easily adapted to any parser output, for any language. The methodology is simple and it is based on the use of regular expressions. This approach brings about a number of advantages (the use of grammars as a corpus of annotated examples; partial or total disambiguation of ambiguous subordinators; information about the position of the subclause relative to the main clause; labelling of unusual constructions; labelling of several subordinate clauses in the same sentence) which encourage further investigations.

1 Introduction

In this paper, we explore the possibility of labelling semantic/adverbial clauses (i.e. *temporal* clauses, *purpose* clauses, *concession* clauses, etc.), complement/nominal clauses (i.e. *ing*-clauses, *that*-clauses, etc.), and complex noun phrases (NPs) using syntactic patterns built on the syntactic annotation returned by a parser. We consider the first two families of clauses as a kind of subordination, therefore we will call them collectively *subordinate* clauses, including under this cover term also complex NPs which are not, properly speaking, subordinate clauses but complex structures (see Biber et al. 1999: 573 ff.).

The need of labelling subordinate clauses derives from our effort to expand the set of features used in automatic genre and text type identification (see Santini 2004b for a review of the concepts of genre and text types together with the automatic approaches suggested so far). However, this kind of annotation can also be useful in many other NLP tasks, such as the analysis of rhetorical/discoursal strategies, authorship attribution or computational stylistics, natural language generation, information extraction, question answering and so on.

By genres and text types we mean, broadly speaking, a classification of documents which is topic-independent. Genres such as *editorials* or *reviews* can be about any topic, for instance editorials can deal with war, politics, ethics, sports, etc.; reviews can comment on films, books, festivals, etc. The same is true for text types such as *argumentation* or *instruction*, which can be used in any discipline or domain. The idea that different "kinds" of documents entail the use of certain syntactic constructions is not new (Biber 1988: 229-230, Baayen et al. 1996, etc.). However, even if syntax is acknowledged to be revealing (although sometimes reluctantly, see Aaronson 1999), it has often been neglected in genre categorization studies, because the extraction of syntactic features is considered to be computationally expensive and time-consuming (Karlgren 2000, Kessler et al. 1997). The 67 linguistic features selected by Biber more than 15 years ago (Biber 1988: 73-75, 221-245) are based mainly on the identification of certain lexical items, even when the features are syntactic, because NLP tools were quite limited at that time. For example, he based the identification of adverbial clauses on the presence of specific subordinators, such as *although* for concessive clauses, and *because* for causative clauses. However, the lexically-based approach to syntax is quite limited, because other subordinators can be ambiguous. To overcome ambiguity,

Biber used only unambiguous subordinators; for example *because* is the only causative subordinator included in his features, being the only one "to function unambiguously as a causative adverbial. Other forms, such as *as*, *for* and *since*, can have a range of functions, including causative" (Biber 1988: 236).

More recently, Part-of-Speech (POS) trigrams have been proposed as "shallow" syntactic features, but they are very corpus-dependent (see Argamon et al. 1998 and Santini 2004a for POS trigrams extraction methods), which means that their exportability to other corpora is not guaranteed.

We suggest that the use of syntactic patterns can represent a useful alternative among other methods for syntax extraction proposed so far. In this paper, we concentrate on the automatic labelling of subordinate clauses. Our approach is parser-dependent, but it can be easily adapted to any parser output, for any language (here the working language is English). The methodology is simple and straightforward, as illustrated in Section 3.

This paper is organized as follows: Section 2 briefly reports on neighbouring experiences with syntactic patterns; Section 3 explains the methodology; Section 4 describes two preliminary attempts to evaluate the approach; Section 5 lists the advantages of the approach together with a number of tasks for future work.

2 Previous work

As far as we know, syntactic patterns of any kind have never been tried for genre and text type identification.

Instead, surface or syntactic patterns are regularly used for information extraction (IE) and question answering (QA), as in Ravichandran et al. 2003, Ravichandran and Hovy 2002, Hovy et al. 2002, Soubbotin and Soubbotin 2001, Riloff 1996, etc. The idea behind the use of these patterns is that certain types of answers and certain types of information are expressed using characteristic phrases. Therefore with questions like "When was X born?", typical answers are: "Mozart was born in 1756", "Gandhi (1869-1948)", etc. Patterns for such answers could be: `NAME was born in BIRTHDATE`, `NAME (BIRTHDATE-` (from Ravichandran and Hovy 2002). For extracting information about bombing, a pattern such as `SUBJ was bombed by PP` would return information such as "World Trade Center was bombed by terrorists" (from AutoSlogTS flowchart, Riloff 1996). The purpose of such patterns is to extract snippets of content by making use of some syntactic components to guarantee a certain degree of generalization.

The aim of the syntactic patterns described in this paper, instead, is to label sentences at subclause level and use this kind of annotation for several different NLP tasks.

3 Methodology

3.1 Use of Grammars as a Corpus of Annotated Examples

To our knowledge, there are no public corpora available with annotation at subclause level, such as *manner clause*, *space clause* and so on. For example, the British National Corpus (BNC) includes only morphological annotation (CLAWS-C5 tagset) and the Penn Treebank II Bracketing allows only extraction of predicate-argument structure. All existing public corpora are important and valuable resources, but creating additional manually-checked annotation on the top of existing annotated corpora is very expensive in terms of time and financial support. Of course, it is also a long-lasting resource, as the positive experience of Carlson et al. (2003) shows.

As it is not always possible to fund human annotators, in our study we propose the use of grammars as a corpus of syntactically annotated examples at subclause level. Annotated examples copied from grammars do not require any further annotation or confirmation by humans, as they represent points

of reference of linguistic knowledge. For this task, we used two comprehensive grammars, Quirk et al. (1985), and Biber et al. (1999) .

3.2 Steps

The approach to the creation of syntactic patterns for labelling subordinate clauses includes the following steps:

1. Copying examples of subordinate clauses from grammars into file(s). At this stage complement/nominal clause and complex NP were copied from Biber et al. (1999), while semantic/adverbial clauses from Quirk et al. (1985).
2. Parsing the file(s) containing the examples of syntactic constructions. The parser for English used here is Connexor by Tapanainen and Järvinen (1997).
3. Tabulation of the parses in a convenient form, more specifically restoring the horizontal alignment from the Connexor vertical output (see Steps 2 and 3 below).
4. Creation of a set of patterns by identifying the common elements of the parses for each syntactic construction and replacing the optional elements with regular expressions.
5. Creation of an algorithm to identify the sets of patterns in running texts.

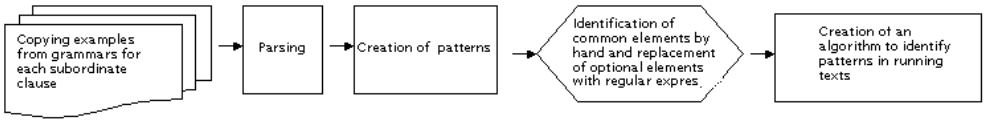

Figure 1. Pipeline for the creation and detection of syntactic patterns.

The five steps will be illustrated using the concessive clause as an example.

Step 1: Grammars as a Corpus of Examples. The Concessive Clause

Examples of clause of concession were copied into a file from Quirk et al. (1985: 1097-1102).

Step 2: Parsing

The file was parsed and a parse was returned for each example.

For instance, the sentence:

```
<!-- : Although he had just joined the company, he was treated exactly
like all the other employees.-->
```

was parsed as follows:

#	Text	baseform	Syntactic relation	Syntax and Morphology
1	Although	although	pm:>5	@CS %CS CS
2	he	he	Subj:>3	@SUBJ %NH PRON PERS NOM SG3
3	had	have	v-ch:>5	@+FAUXV %AUX V PAST
4	just	just	meta:>5	@ADVL %EH ADV
5	joined	join	cnd:>11	@-FMAINV %VA EN
6	the	the	det:>7	@DN> %>N DET
7	company	company	obj:>5	@OBJ %NH N NOM
8	,	,		

9	he	he	subj:>10	@SUBJ %NH PRON PERS NOM SG3
10	was	be	v-ch:>11	@+FAUXV %AUX V PAST
11	treated	treat	main:>0	@-FMAINV %VP EN
12	exactly	exactly	man:>11	@ADVL %EH ADV
13	like	like		@ADVL %EH PREP
14	all	all	det:>15	@DN> %>N DET
15	the	the	det:>16	@DN> %>N DET
16	other	other	attr:>17	@DN> %>N DET
17	employees	employee		@NH %NH N NOM PL
18	.	.		
19	<s>	<s>		

Figure 1. A sentence parsed by Connexor.

Step 3: Pattern Creation

Connexor returns several types of annotation[1]. For the creation of syntactic patterns, the priority was given to syntax, i.e. the annotation starting with an **@** sign, occasionally integrated by other types of annotation, for example the syntactic relation **main:>** that identifies the main verb in the whole sentence, or lexical items, such as **although**, to represent subordinators. Wrong parses were not used to build patterns. A simple algorithm based on string manipulation clears out all unnecessary information and returns the following pattern from the parse shown in Figure 1:

```
<!-- : Although he had just joined the company, he was treated exactly
like all the other employees.-->
```

```
although@CS @SUBJ have@FAUXV @ADVL FMAINV_EN @DN> @OBJ , @SUBJ be@FAUXV
main_FMAINV_EN @ADVL @ADVL @DN> @DN> @DN> @NH
```

Reading such a pattern is very easy:

```
although@CS          = 'although' in the role of subordinate conjunction
@SUBJ                = subject of the clause
have@FAUXV           = 'have' as auxiliary
@ADVL                = adverb
FMAINV_EN            = past participle
@DN>                 = determiner
@OBJ                 = object
be@FAUXV             = 'be' as auxiliary
main_FMAINV          = main verb in the sentence
@ADVL                = adverb
@NH                  = noun head
```

Step 4: Manual Identification of Common Elements across Patterns

What are the common elements between the two following patterns?

[1] The complete annotation scheme is available online at http://www.connexor.com/demo/doc/enfdg3-tags.html

Sentence 1:

```
<!-- : Although he had just joined the company, he was treated exactly like all the other employees.-->
```

```
although@CS @SUBJ have@FAUXV @ADVL FMAINV_EN @DN> @OBJ , @SUBJ be@FAUXV main_FMAINV_EN @ADVL @ADVL @DN> @DN> @DN> @NH
```

Sentence 2:

```
<!-- : Although Sam had told the children a bedtime story, June told them one too. -->
```

```
although@CS @SUBJ have@FAUXV FMAINV_EN @DN> @I-OBJ @DN> @A> @OBJ , @SUBJ main_FMAINV_PAST @I-OBJ @OBJ @ADVL
```

It's very easy to detect their similarity: both start with *although*, both have a subject followed by a verb in the subordinate clause, both have a subject and a main verb in the matrix clause. The common elements are:

```
although@CS @SUBJ FMAINV @SUBJ main_FMAINV
```

In order to make this pattern flexible to any number of optional elements occurring between each component, we can simply use regular expressions, provided by many programming languages (the use of regular expression is well-explained in Friedl 1997). The use of a 'non-greedy' quantifier (i.e. a quantifier that finds the minimum number of character matching the pattern) improves efficiency. In many cases, the metacharacters `*?` are enough to meet the need of flexibility and efficiency. The pattern filled in by regular expressions has the following form:

```
although@CS.*?@SUBJ.*?FMAINV.*?@SUBJ.*?main_FMAINV
```

and matches both examples shown above.

It is worth highlighting that the added value of a syntactic pattern for an unambiguous subordinator such as *although* is the additional information given by the *position of the subclause relative to the main clause*. In fact, the position of the subordinate clause in the sentence is considered to be genre/register connected and also influenced by coherence and information structuring (cf. Biber et al. 1999: 830-838). The unmarked choice is the final position for all subclause types, therefore a variation of this choice can be informative. For example, semantic/adverbial clauses tend to be in initial position when they contain "given" information referring to previous discourse, while the main clause presents "new" information; on the contrary, when the main clause bears the "given" information, the semantic/adverbial clauses tend to be in final position (see Biber et al. 1999: 835-836).

More importantly, syntactic patterns can help disambiguate ambiguous subordinators, for example *though*. *Though* can be a subordinator and a linking adverbial (Biber et al. 1999:850). With the use of syntactic patterns, the two roles cannot be confused. For example, in the following sentence:

```
No goals were scored, though it was an exciting game.
@DN> @SUBJ @FAUXV main_FMAINV , though@CS @SUBJ FMAINV @DN> @A> @PCOMPL-S
```

the subclause will be labelled unambiguously as concessive clause, while:

```
He went, though.
@SUBJ main_FMAINV , though@ADVL
```

will not. Even if the parser makes a tagging mistake, the pattern for *though* subordinator ensures the full disambiguation.

Step 5: Labelling Algorithm

The procedure of finding common elements across patterns was repeated for all examples of subordinate clauses included in this study (see Subsection 3.3). For each subclause, a set of patterns was built. Each set of patterns was searched and labelled by a subroutine, as in the snippet below:

```
sub labelling_concession_clause
{
undef @initial;
undef @final;
undef @special;

@initial =
(
"although\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",
"_though\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",

"although\@CS.*?\@NH , .*?$s_v_main",
"_though\@CS.*?\@NH , .*?$s_v_main",

"even though\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",
"even though\@CS.*?NH.*?FMAINV.*?$s_v_main",
"even though\@CS.*?\@FMAINV_EN.*?$s_v_main",

"even if\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",
"even if\@CS.*?NH.*?FMAINV.*?$s_v_main",
);

    for ($count_features = 0; $count_features<@initial; $count_features++)
    {
        if(/(@initial[$count_features])/)

            {
                print $1;
                print "\t ";
                print "---+> concession_clause_initial\n";
            }
    }
```

Figure 2. Example of a subroutine written in Perl to label syntactic patterns of concessive clause.

3.3 Coverage

The subordinate clauses partially or fully covered by syntactic patterns include ten semantic/adverbial clauses, nine complement/nominal clauses, and the complex NP.

The semantic/adverbial clauses are the following:

1. concession clause (initial, final, special)
2. conditional clause (initial, final, special)
3. contrast clause
4. exception clause
5. purpose clause
6. reason clause (initial, final)
7. result clause
8. similarity manner comparison clause
9. space clause (initial, final)
10. time clause (initial, final, incidental, instructional)

The labels “initial”, “final”, “incidental”, “special” and “instructional” indicate respectively initial position relative to the main clause, final position, incidental position, a special or unusual syntactic construction and, finally, a construction used mainly in instructional texts.

Here is the list of nine complement/nominal clauses, plus the complex NP:

1. verb+that clause
2. adjective+that clause
3. that omission
4. wh-clause
5. verb+to clause
6. adjective+to clause
7. verb+ing clause
8. comparative clause
9. relative clause
10. complex NP

The full list of syntactic patterns, with examples, created so far is in Santini (2005:18-37). It is important to point out that these syntactic patterns do not cover all the possible patterns for the subclauses listed above, but only a number of them. Moreover, six subordinators are highly ambiguous (*as*, *if*, *since*, *when*, *whereas*, *while*)[2], therefore for these subordinators only those syntactic patterns which could unambiguously represent a subclause were included. For example, *if* can be both conditional and concessive, but the pattern *if* + past participle, as in the sentence *The grass will grow more quickly if watered regularly* (Santini 2005:27), can never be concessive, so this pattern is unambiguously labelled as conditional. Also the position on the subclause in the sentence can contribute to the disambiguation of subordinators. In the case of *as*, which can mark both temporal and similarity subclauses, the pattern *as* in initial position, followed by a main clause starting with *so*, as in the sentence *As the moth is attracted by a light, so he was attracted by her* (Santini 2005:30), will be unambiguously labelled as a similarity subclause.

Even though these subordinators have not been entirely disambiguated, and the coverage of the syntactic patterns in general is still incomplete, these examples give a flavour of the power and the potential of such an approach. So far, for genre and text type identification, the tendency has been to use only subordinators that are unambiguous, such as *although* for concessive clause or *because* for reason clause (see Introduction), or to overextend the distribution of a subordinator (for example, *if* always conditional, or *when* always temporal). The use of patterns help overcome these limitations.

4 Evaluation

4.1 Initial Assessment of Automatic Labelling

As mentioned earlier, there are no corpora available in English with syntactic annotation at subclause level. Therefore the accuracy of automatic labelling (Step 5 above) cannot be measured against any benchmark. In order to get a rough idea of how well the automatic labelling performs, the syntactic patterns of five subclauses (*time, manner, purpose, concession,* and *reason*) created with examples taken from Quirk et al. (1985) were tested against 34 sentences containing the same subclauses taken from Biber et al. (1999).

Figure 4 shows a snippet of the output of the labelling algorithm and is useful to highlight some problems connected to a thorough evaluation of this approach.

---

[2] *As* is a subordinator for reason (ex.: *As Jane was the eldest, she looked after the others*), similarity (ex.: *She cooks a turkey as her mother did*) and time (ex.: *As I drove away, I saw her*); *if* for concession (ex.: *If he is poor, he's honest*) and condition (ex.: *If you put the baby down, she will scream*); *since* for reason (ex.: *Since we live near the sea, we often go sailing*) and time (ex.: *Since I last saw her, she has dyed her hair*); *when* for concession (ex.: *She paid, when she could have entered free*), time (ex.: *When I last saw you, you lived in Washington*) and space (ex.: *Take the right fork when the road splits into two*); *whereas* for concession (ex.: *Whereas the amendment was supported in the Senate, its fate is doubtful*) and contrast (ex.: *I ignore them, whereas my husband is worried of what they think of us*); *while* for concession (ex.: *While he has many friends, Peter is lonely*), contrast (ex.: *John teaches physics, while Mary teaches chemistry*) and time (ex.: *He cut himself while shaving*). All the examples come from Quirk et al. (1985).

The output shows different kind of information. For instance, Sentence 1 was copied from Biber et al. 1999 (called LGSWE in the output) on page 818, from examples provided for time adverbials. This information was enriched by the author of this paper with the addition of the position of the subclause relative to the main clause, and the subordinator employed (`<!-- : ****** time_clause_initial_when ****** -->`). Sentence 1 is automatically labelled as *time_clause_initial* and *verb_to_clause*, both labels are correct. One limitation with the use of grammars for evaluation purposes is that it allows the evaluation of only one label per sentence, because the emphasis in grammars is mostly on a single linguistic phenomenon at time. Therefore, the label *verb_to_clause* (`expects to recover`) cannot be objectively evaluated.

As for the other sentences, Sentence 2 receives only the label *complex NP* (`good thoughts`) because the pattern with *until* followed by *be* is not available yet. Sentence 3 is labelled correctly as *time_clause_initial*, and finally Sentence 4 do not receive any label, because, as mentioned before, *as* is a highly ambiguous subordinators, which has not be thoroughly disambiguated yet.

```
<!-- : ====================== TIME CLAUSES LGSWE ====================== -->
Sentence 1:
<!-- : time_clauses_LGSWE_page_818 - When the units are sold, the city expects
to recover all but its $825,000 initial investment.  --><!-- : ******
time_clause_initial_when ****** -->
Pattern 1:
when_tmp_@ADVL__ADV @DN> @SUBJ be@FAUXV_V sell_tmp_@FMAINV__EN , @DN> @SUBJ
main_FMAINV  ---+> time_clause_initial
Pattern 2:
factual_verb_main_FMAINV_SG3 @INFMARK> FMAINV    ---+> verb_to_clause


Sentence 2:
<!-- : time_clauses_LGSWE_page_822 - These good thoughts carry me until I'm
right downtown.   --><!-- : ****** time_clause_final_until ****** -->
Pattern 1:
@A> @SUBJ    ---+> complex_np


Sentence 3:
<!-- : time_clauses_LGSWE_page_822 - When it would not suck from the bottle she
fed it with a teaspoon, fretting with impatience when it coughed and sputtered
and cried.  --><!-- : ****** time_clause_initial_when ****** -->
Pattern 1:
when_tmp_@ADVL__ADV _it@SUBJ_PRON @FAUXV not_neg_@ADVL__NEG suck_tmp_@FMAINV__V
from_sou_@ADVL__PREP @DN> @<P @SUBJ main_FMAINV ---+> time_clause_initial


Sentence 4:
<!-- : time_clauses_LGSWE_page_822 - Tom could see her in tears as she wrote
it.--><!-- : ****** time_clause_final_as ****** -->


no patterns available yet
[...]
```

Figure 4. A snippet from the output returned by the labelling algorithm.

Three categories were used to give an idea of the quality of the automatic labelling: **C(orrect), I(ncorrect), NAY (not available yet)**. Out of 34 sentences, 21 were labelled correctly, 2 received incorrect labels, and for 11 sentences the patterns were not yet available. If we consider that the patterns for these subclauses were built using only a limited number of examples from Quirk et al. (1985), these results show the good potential of the approach. Presumably, adding new patterns from different grammars or from corpora annotated at subclause level, when they will be available, will extend the coverage of the automatic labelling.

As for the errors performed by the parser, errors were detected especially with special constructions, such as the concessive clause: *Fool that he was he managed to evade his pursuers.*

#	Text	baseform	Syntactic relation	Syntax and Morphology
1	Fool	fool	main:>0	@+FMAINV %VA V IMP
2	that	that	pm:>4	@CS %CS CS
3	he	he	subj:>4	@SUBJ %NH PRON PERS NOM SG3
4	was	be	obj:>1	@+FMAINV %VA V PAST
5	,	,		
6	he	he	subj:>7	@SUBJ %NH PRON PERS NOM SG3
7	managed	manage		@+FMAINV %VA V PAST
8	to	to	pm:>9	@INFMARK> %AUX INFMARK>
9	evade	evade	obj:>7	@-FMAINV %VA V INF
10	his	he	attr:>11	@A> %>N PRON PERS GEN SG3
11	pursuers	pursuer	obj:>9	@OBJ %NH N NOM PL
12	.	.		
13	<p>	<p>		

Figure 5. Wrong parse for a special construction, concessive clause (Quirk et al. 1985:1098).

As you can see in Figure 5, the adjective *fool* is parsed as an imperative. As stated above, patterns are not built from examples parsed incorrectly. The impact of errors performed by the parser will be clear only when a comprehensive benchmark is devised to evaluate the automatic labelling of subordinate clauses.

4.2 Syntactic Patterns as Features for Web Genre Classification

It is important to point out again that syntactic patterns were devised primarily as features for genre and text type identification. To test these features for genre classification, we built two datasets for three web genres: *blogs, FAQs,* and *frontpages* (200 documents per web genre). The first dataset was built with normalized frequency counts of syntactic patterns of the following subclauses:

1. adjective_that_clause
2. adjective_to_clause
3. comparative_clause
4. complex_np
5. concession_clause_final
6. concession_clause_initial
7. concession_clause_special
8. conditional_clause_final
9. conditional_clause_initial
10. conditional_clause_special
11. contrast_clause
12. exception_clause
13. purpose_clause
14. reason_clause_final
15. reason_clause_initial
16. relative_clause
17. result_clause
18. similarity_manner_comp_cl.
19. space_clause_final
20. space_clause_initial
21. that_omission
22. time_clause_final
23. time_clause_incidental
24. time_clause_initial
25. time_clause_instructional
26. verb_ing_clause
27. verb_that_clause
28. verb_to_clause
29. wh_clause

The second dataset was build with normalized frequency counts of the following subordinators:

1.	although	15.	if	29.	that
2.	as	16.	immediately	30.	though
3.	as_if	17.	in_case	31.	unless
4.	as_long_as	18.	in_order_to	32.	until
5.	as_soon_as	19.	just_so	33.	what
6.	as_though	20.	never	34.	when
7.	because	21.	once	35.	whenever
8.	before	22.	only	36.	where
9.	but_for	23.	save_that	37.	whereas
10.	but_that	24.	since	38.	wherever
11.	even_if	25.	so	39.	which
12.	even_though	26.	so_long_as	40.	while
13.	except	27.	so_long_to	41.	whoever
14.	how	28.	so_that	42.	why

The learning algorithm used for this task was an SVM classifier as implemented in Weka (Witten and Frank 2000). A 10-fold cross-validation was run 10 times with different seeds, and the accuracy results were averaged. The dataset containing 29 syntactic patterns returned an averaged accuracy of 86.2%, while the dataset with the 42 subordinators reached an averaged accuracy of 83.6%. The gap of 2.6% in the accuracy is quite important if we consider that the list of syntactic patterns is still incomplete, while the list of subordinators is more comprehensive. Presumably, with a better coverage of syntactic patterns the classification accuracy will increase. If so, labelling subordinating clauses appears to be more profitable than using lexical items.

5 Conclusions and Future Work

The approach proposed in this paper brings about a number of advantages, for example:

1) the use of grammars as a corpus of annotated examples at subclause level (Subsection 3.1);

2) partial disambiguation of some ambiguous subordinators, such as *as* and *if* (Subsection 3.3);

3) disambiguation of different syntactic roles (see the example of *though*, Step 4 in Subsection 3.2);

4) information about the position of the subclause (initial, final, etc.) relative to the main clause, which can be important for some NLP tasks, such as genre and text type identification, or for a better understanding of information structuring (Step 4 in Subsection 3.2).

5) Unusual constructions, like the following concessive clause (Quirk et al. 1985:1098), can be easily identified, without creating any noise:

```
Naked as I was, I braved the storm.
@PCOMPL-S as@CS @SUBJ be_FMAINV , @SUBJ main_FMAINV
```

In this pattern, there is a subject complement, followed by *as* in the role of conjunction, followed by a subject, followed by a verb *be*, followed by another subject, followed by a main verb. This special construction will be labelled unambiguously as concessive clause.

6) Several subordinate clauses can be identified and labelled in a single sentence. For example:

```
Although Sam had told the children a bedtime story, June told them one
too, because she was eager to see their reaction.
```

both patterns:

```
although@CS @SUBJ.*?FMAINV.*? , @SUBJ.*?main_FMAINV
```

and

```
SUBJ.*?FMAINV.*?because@CS.*?SUBJ.*?FMAINV
```

are applicable, therefore the sentence is fully labelled as bearing a concessive clause in initial position and a reason clause in final position.

All these benefits are encouraging, but lots remain to be done. First of all, a more reliable way to evaluate the automatic labelling must be worked out. As shown in Section 4, each sentence might have several labels. Are human annotators the only solution for this problem? Discussion and suggestions on this specific point are necessary and welcomed. Second, the list of syntactic patterns must be enlarged for the initial set of subclauses used in this study, but also patterns for other subclauses or complex structures must be built. Last but not least, more effort must be put on the full disambiguation of ambiguous subordinators. This is not an easy task, but the linearity of syntactic patterns with the integration of additional linguistic information returned by the parser can help substantially.

6 References

Aaronson S. (1999), *Stylometric Clustering: A comparative analysis of data-driven and syntactic features*, Project report available at http://www.cs.berkeley.edu/~aaronson/sc/report.doc

Argamon S., Koppel M. and Avneri G. (1998), Routing documents according to style, *Proceedings of the First International Workshop on Innovative Internet Information Systems* (IIIS-98).

Baayen H., Halteren (van) H. and Tweedie F. (1996), Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, *Literary and Linguistic Computing*, 11.

Biber, D. (1988), *Variations across speech and writing*, Cambridge University Press, Cambridge.

Biber, D. (1989), A typology of English texts, *Linguistics*, Vol. 27, 3-43.

Biber D., Johansson S., Leech G., Conrad S. and Finegan E. (1999), *Longman Grammar of Spoken and Written English, Longman*, Harlow.

Carlson L., Marcu M. and Okurowski M. E. (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, in J. van Kuppevelt and R. Smith (eds.) *Current Directions in Discourse and Dialogue*, (Kluwer Academic Publishers), 85-112.

Friedl J. (1997), *Mastering Regular Expressions*, O' Reilly, Beijing, Cambridge.

Hovy E., Hermjakob U. and Ravichandran D. (2002), A Question/Answer Typology with Surface Text Patterns, *Proceedings of the DARPA Human Language Technology Conference* (HLT), San Diego, CA.

Karlgren J. (2000), *Stylistic Experiments for Information Retrieval*, Thesis submitted for the degree of Doctor of Philosophy, Department of Linguistics, Stockholm University.

Kessler B., Numberg G. and Shütze H. (1997), Automatic Detection of Text Genre, Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.

Quirk R., Greenbaum S., Leech G. and Svartvik J. (1985), *A Comprehensive Grammar of the English Language*, Longman.

Riloff E. (1996), Automatically Generating Extraction Patterns from Untagged Texts, *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (AAAI-96).

Santini M. (2004a), A Shallow Approach To Syntactic Feature Extraction For Genre Classification, *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics* (CLUK 04), University of Birmingham (UK), 6-7 January, 2004.

Santini M. (2004b), *State-of-the-art on Automatic Genre Identification*, Technical Report ITRI-04-03, 2004, ITRI, University of Brighton (UK).

Santini M. (2005), *Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features*, Technical Report ITRI-05-02, 2004, ITRI, University of Brighton (UK).

Soubbotin M. and Soubbotin S. (2001), Patterns of Potential Answer Expressions as Clues to the Right Answer, *Proceedings of the TREC-10 Conference*, NIST, Gaithersburg, MD.

Ravichandran D. and Hovy E. (2002), Learning Surface Text Patterns for a Question Answering system, *Proceedings of the 40th ACL conference*, Philadelphia, PA.

Ravichandran D., Ittycheriah A. and Roukos S. (2003), Automatic Derivation of Surface Text Patterns for a Maximum Entropy Based Question Answering System, *Proceedings of the HLT-NAACL conference*, Edmonton, Canada.

Tapanainen P., Järvinen T. (1997), A non-projective dependency parser, *Proceedings of the 5th Conference on Applied Natural Language Processing*.

Witten I. and Frank E. (2000), *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco.

Agreement Patterns in Corpora*

Aline Villavicencio and Louisa Sadler
Department of Language and Linguistics
University of Essex
Wivenhoe Park
Colchester, CO4 3SQ, UK
{avill, louisa}@essex.ac.uk

Abstract

Syntactically annotated corpora are very valuable resources that can be used to provide crucial evidence for the occurrence of particular linguistic constructions in a given language. In the case of agreement processes, the analysis of the different strategies found in a language with a rich agreement profile is paramount for testing the limits of current theories of agreement, and the availability of syntactically annotated corpora enables such cases to be unearthed. This paper discusses agreement patterns of postnominal adjectives in Portuguese, with evidence for different strategies gathered from corpora. We focus on cases of closest conjunct agreement in NP (and noun) coordinations. The results obtained are used to clarify the conditions under which agreement with the closest conjunct is grammatical and discuss the implications of these findings for an HPSG analysis of agreement.

1 Introduction

Syntactically annotated corpora are very valuable resources that can be used to provide crucial evidence for the occurrence of particular linguistic constructions in a given language. In the case of agreement processes, the availability of such corpora enables the investigation of the different strategies used in a given language. This data can subsequently be used, for example, for outlining a theory of agreement. In this paper we discuss how an annotated corpus can be used to investigate agreement patterns and the frequency with which they occur in data, helping to understand the conditions under which some of these patterns are used.

Agreement phenomena in general have received considerable attention in recent years from the linguistic community (e.g. Corbett (1991); Pollard and Sag (1994); Sadler (1999); Kathol (1999); Dalrymple and Kaplan (2000); Wechsler and Zlatić (2003); King and Dalrymple (2004), among others). Coordinate structures in particular present a challenging picture as they allow agreement patterns that are not found in non-coordinate structures, as coordinated nouns may jointly control agreement with dependents such as adjectives. For instance in the following Portuguese sentence, a masculine singular noun is coordinated with a feminine singular noun, and the coordinate structure has masculine plural agreement with the postnominal adjective:

*The research reported on here was carried out as part of a larger project on Noun Phrase Agreement and Coordination funded by the AHRB (RG/APN17606/AN109, to Sadler and Dalrymple), whose support we gratefully acknowledge.

(1) Ele atribuía o erro e a inconstância humanos aos caprichos da
He attributed the.MSG error.MSG and the.FSG inconstancy.FSG human.MPL to the caprices of
experiência
experience

'He attributed the human error and inconstancy to the caprices of experience'

The analysis of the different agreement strategies found in a language with a rich agreement profile is paramount for defining the characteristics that a theory of agreement should have. In this paper we investigate agreement patterns between postnominal adjectives and nouns found in Portuguese NP (or N) coordinations, with evidence for different strategies as well as for their frequency gathered from corpora. We concentrate on agreement of gender and number between postnominal adjectives and the coordinated NPs, in particular closest conjunct agreement. The picture of agreement patterns in Portuguese NPs which emerges from the corpus study is a complex one and the results obtained are used for offering some observations on the conditions under which agreement with the closest conjunct is grammatical. This paper starts with an overview of agreement patterns between postnominal adjectives and coordinated NPs in Portuguese, in section 2, and a discussion of the findings gathered in the corpus study, in section 3. A discussion of the implications of the corpus study for a theory of coordination is presented in section 4, followed by the conclusions and future work.

2 Agreement Patterns in Portuguese

The current paper reports on part of an ongoing study of agreement within coordinate NPs in Portuguese, and focuses in particular on the agreement behaviour of postnominal adjectives in coordinate NPs, where the postnominal adjectives scope over the coordinate NPs as a whole. In non-coordinate structures, Portuguese postnominal adjectives agree straightforwardly in number and gender with the nouns they modify, as illustrated in the examples below:

(2) a parede colorida/vermelha/*pintadas/*colorido/*vermelho/*pintados
the.FSG wall.FSG coloured.FSG/red.FSG/*painted.FPL/*coloured.MSG/*red.MSG/*painted.MPL

(3) o.MSG teto.MSG *colorida/*vermelha/*pintadas/colorido/vermelho/*pintados
the ceiling *coloured.FSG/*red.FSG/*painted.FPL/coloured.MSG/red.MSG/*painted.MPL

Coordinate structures on the other hand present a much wider range of agreement patterns, since coordinated nouns often jointly control agreement on determiners, adjectives and other dependents within the NP. A strategy common to many languages (and widely discussed in the literature) involves a familiar type of syntactic resolution of agreement features. The following examples illustrate **Resolution Agreement** with coordinations of both same gender (see (4) and (5)) and different gender nouns (see (6)).

(4) a banana e a pera maduras
the.FSG banana.FSG and the.FSG pear.FSG ripe.FPL

(5) o carro e o barco novos
the.MSG car.MSG and the.MSG boat.MSG new.MPL

(6) o homem e a mulher modernos
the.MSG man.MSG and the.FSG woman.FSG modern.MPL
the modern man and woman

Resolution Agreement in Portuguese involves semantic number resolution (to plural in the general case, excluding examples of single entity coordinations such as “my friend and colleague”) and resolution to the masculine for gender, a widespread strategy in the Romance languages and beyond:

(7) If all conjuncts are GEN = FEM, resolve to FEM
else, resolve to MASC

While syntactic resolution is a very widespread strategy, a second strategy which is well attested crosslinguistically involves a form of single conjunct agreement, namely **Closest Conjunct Agreement** (CCA). This strategy is also found in Portuguese and is exemplified by (8). In this example, syntactic resolution of both number and gender are suspended and the postnominal adjective bears agreement features coding the closest (i.e., final) coordinate noun.

(8) estudos e profissão monástica
studies.MSG and profession.FSG monastic.FSG

Note that some examples of closest conjunct agreement involve cases in which the conjuncts are synonyms (9) or are part of an enumeration (10):

(9) As maldições se cumpriam no povo e gente hebreia
The curses themselves fell in the.MSG people.MSG and persons.FSG hebrew.FSG

(10) No cumprimento de seus deveres tinha aquele homem um zelo, uma
In the fulfillment of his obligations had that man a.MSG zeal.MSG, a.FSG
severidade, uma exação extraordinária
severity.FSG, an.FSG exactness.FSG extraordinary.FSG

Although the existence of closest conjunct agreement within Portuguese coordinate NPs has not received attention in the theoretical linguistic literature and beyond, at least one detailed descriptive grammar of Portuguese does provide some discussion and exemplification of this phenomenon. The agreement possibilities discussed by Torres (1981) for postnominal adjectives with coordinate NPs are spelled out in table 1, where **NP1, NP2** and **Adj** refer to the number and gender of the first conjunct, second conjunct, and adjective, respectively.

Table 1: Summary of Agreement Strategies in Portuguese

	Strategy	NP1	NP2	Adj
1	Resolved(G,N)	MSG	FSG	MPL
2	CCA(G,N)	MSG	FSG	FSG
3	Resolved(G,N) (*)	MSG	FPL	MPL
4	CCA(G,N) (*)	MSG	FPL	FPL
5	Resolved(G,N)	MPL	FSG	MPL
6	CCA(G,N)	MPL	FSG	FSG
7	Resolved(G,N) (*)	MPL	FPL	MPL
8	CCA(G,N) (*)	MPL	FPL	FPL
9	Resolved(G,N) (*)	FSG	MSG	MPL
10	CCA(G,N) (*)	FSG	MSG	MSG
11	Resolved(G,N) or CCA(G,N) (*)	FSG	MPL	MPL
12	Resolved(G,N) (*)	FPL	MSG	MPL
13	CCA(G,N) (*)	FPL	MSG	MSG
14	Resolved(G,N) or CCA(G,N) (*)	FPL	MPL	MPL

As is evident from the table above, Torres assumes the existence of two patterns in Portuguese - CCA (in gender and number) and resolution (of gender and number). Note however that *in principle* the rows marked with an asterisk could be interpreted as displaying a "mixed" strategy. That is, given that a language permits the CCA pattern for both number and gender shown above, there are in principle two further closest conjunct agreement patterns which might operate. These are patterns in which the agreement features of number and gender "come apart", that is cases in which gender agreement is with the closest conjunct while number is (semantically) resolved, and cases in which gender is (syntactically) resolved while number marking reflects the number value of the closest conjunct. Of course for the range of data that Torres gives, there is no reason to further hypothesize "mixed" controllers in this way, given that the two "simple" patterns of CCA and resolution cover the data. However, we note that there is positive existence for the first of these in Portuguese, evidence which is not discussed in Torres:

(11) o sofrimento e a experiência vividas
the.MSG suffering.MSG and the.FSG experience.FSG lived.FPL

This is a clear case in which the postnominal adjective scopes over the NP coordination as a whole while the feminine gender on the adjective indicates gender agreement with the closest conjunct yet plurality on the adjective indicates a resolved feature, since each NP is actually singular. This strategy is shown in table 2.

Table 2: Further Agreement Strategies - I

	Strategy	NP1	NP2	Adj
15	CCA(G), Resolved(N)	MSG	FSG	FPL
16	CCA(G), Resolved(N)	MPL	FSG	FPL

The existence of this possibility is also noted by Camacho (2003) for Spanish NP/N coordinations, where CCA of gender is again combined with resolution of number - below we give further Portuguese examples preceded by an example for Spanish.

(12) Ejerce influencia en el crecimiento y la reproducción genéticas
Exercises influence in the.MSG growth.MSG and the.FSG reproduction.FSG genetic.FPL

(13) ... para um país com fome de capitais e tecnologia externas
... to a country with hunger for capital.MPL and technology.FSG external.FPL
To a country in need of external capital and technology

(14) ... uma relação entre sobrecarga do organismo e envelhecimento e morte
... a relation between overload of the organism and aging.MSG and death.FSG
prematuras
premature.FPL
A relation between overload of the organism and the premature aging and death.

(15) ... tendo um conhecimento e uma experiência acumuladas que
... having a.MSG knowledge.MSG and an.FSG experience.FSG accumulated.FPL that
permitem...
allow...
http://www.jorplast.com.br/secoes/Jul98.htm

The second possibility involves the options listed in table 3, with either gender resolution and number CCA or furthest conjunct agreement. However, these patterns are not possible for Portuguese, as illustrated by sentence 16, or for Spanish (Camacho, 2003), sentence 17.

Table 3: Further Agreement Strategies - II

	Strategy	NP1	NP2	Adj
17	Resolved(G) and CCA(N)	MSG	FSG	MSG
18	Resolved(G) and CCA(N)	MPL	FSG	MSG

(16) *O currículo e a pesquisa universitário foram discutidos em ...
The.MSG program.MSG and the.FSG research.FSG university.MSG were discussed in ...

(17) *Ejerce influencia en el crecimiento y la reproducción genético
Exercises influence in the.MSG growth.MSG and the.FSG reproduction.FSG genetic.MSG

From this investigation, a complex picture emerges, which is further investigated in the corpus analysis of NP internal agreement patterns in Portuguese, discussed in the next section.

3 A Corpus Investigation

To estimate the frequency with which these agreement strategies are used in coordinate nouns modified by postnominal adjectives, an investigation using an annotated corpus was performed. Of particular interest are cases that employ closest conjunct agreement.

For this analysis we searched the 32,091,996 word NILC/São Carlos corpus (available from http://www.linguateca.pt/) for occurrences of coordinated NPs/Ns modified by postnominal adjectives. This corpus contains Brazilian texts from newspapers, books and essays, among others. The searches specified concordances such as:

```
[pos="DET_artd"] [pos="N" & gen="M" & pessnum="P"]
[pos="KC.*" & word="e"]  [pos="DET_artd"]
[pos="N" & gen="F" & pessnum="P"]
[pos="ADJ" & pessnum="P" & gen="F" ]
```

for the coordination of a Masculine Plural NP (determiner and noun) and a Feminine Plural NP using the conjunction *e* (*and*) postmodified by a Feminine Plural Adjective, and

```
[pos="N" & gen="M" & pessnum="P"] [pos="KC.*" & word="e"]
[pos="N" & gen="F" & pessnum="P"]
[pos="ADJ" & gen="F" & pessnum="P"  ]
```

for the coordination of a Masculine Plural Noun and a Feminine Plural Noun using the conjunction *e* (*and*) postmodified by a Feminine Plural Adjective.

Searches for NP and N concordances such as these were done for each of the combinations of gender and number shown in tables 1, 2 and 3 where the second conjunct is feminine, since we want to focus on the cases where we can unambiguously detect CCA of gender. The other cases are ambiguous between a strategy of resolution to masculine, or a strategy of CCA with the masculine noun.

For NP coordinations a subset of the NILC corpus containing 305 sentences was obtained and for N coordinations a subset with 2,337 sentences. These sentences were manually post-processed so that any cases that involved adjectives that were common to both genders were removed. Only adjectives that overtly reflect gender distinction were kept, as we wanted to test the correlation between the gender of each of the conjuncts and the gender of the adjective. Sentences where the adjective scoped over only one of the conjuncts were also removed. As a consequence, 41 out of the 305 sentences with NP coordinations remained, and 374 out of 2,337 with N coordinations.

These sentences are distributed as shown in table 4 for coordinations of NPs, and table 5 for Ns, where **Initial Frequency** indicates the number of sentences found for the searches before post-processing and **Final Frequency** after post-processing. **Animacy** indicates whether the coordination included animate nouns, **Enumeration**, whether the nouns are part of an enumeration, **Synonyms**, if they are synonyms, and **Other**, if they include cases that are neither enumerations or synonyms. In the results for these aspects, the **Yes** value indicates a pattern that was found in the analysed data and **No** one that was not found.[1]

In terms of number agreement, cases 1, 5, 15 and 16 signal the adoption of resolution to plural as the second conjunct is singular and the adjective is in plural form. Cases 2 and 6 are unambiguously of CCA where the adjective follows the number of the second conjunct. All other cases are ambiguous between a resolution to plural and a CCA strategy.

For gender, cases 1, 3, 5 and 7 adopt a resolution to masculine strategy, while cases 2, 4, 6, 8, 15 and 16 show CCA of gender with the last conjunct. One interesting point to observe is that CCA of gender seems to be frequently employed when compared to resolution to masculine (e.g. compare cases 1 and 2, and 7 and 8).

Some positive evidence was found in the corpus for the patterns that mix CCA of gender and resolution of number (15 and 16), suggesting that Portuguese, like Spanish, allows mixed controllers for gender and number. On the other hand, the lack of evidence for cases 17 and 18 in the corpus data suggests that even if mixed controllers were allowed in Portuguese, there may be some constraints on the acceptable combinations of number and gender.

[1] **No** in particular does not indicate if a pattern is impossible, but only that given the data available, it has not been found.

Table 4: Frequency of Agreement Strategies in Portuguese NP Coordination

	NP1	NP2	Adj	Initial Frequency	Final Frequency	Animacy	Enumeration	Synonyms	Other
1	MSG	FSG	MPL	5	4	Yes	Yes	No	Yes
2	MSG	FSG	FSG	166	13	No	Yes	Yes	Yes
3	MSG	FPL	MPL	0	0				
4	MSG	FPL	FPL	54	7	Yes	Yes	Yes	Yes
5	MPL	FSG	MPL	0	0				
6	MPL	FSG	FSG	0	0				
7	MPL	FPL	MPL	1	1	Yes	No	No	Yes
8	MPL	FPL	FPL	67	15	Yes	Yes	Yes	Yes
15	MSG	FSG	FPL	11	1	Yes	No	No	Yes
16	MPL	FSG	FPL	1	0				
17	MSG	FSG	MSG	2	0				
18	MPL	FSG	MSG	0	0				

Table 5: Frequency of Agreement Strategies in Portuguese N Coordination

	N1	N2	Adj	Initial Frequency	Final Frequency	Animacy	Enumeration	Synonyms	Other
1	MSG	FSG	MPL	32	30	Yes	Yes	Yes	Yes
2	MSG	FSG	FSG	574	37	No	Yes	Yes	Yes
3	MSG	FPL	MPL	3	2	No	Yes	No	Yes
4	MSG	FPL	FPL	264	7	No	Yes	Yes	Yes
5	MPL	FSG	MPL	7	7	No	Yes	No	Yes
6	MPL	FSG	FSG	362	8	No	No	Yes	Yes
7	MPL	FPL	MPL	86	78	Yes	Yes	Yes	Yes
8	MPL	FPL	FPL	957	199	Yes	Yes	Yes	Yes
15	MSG	FSG	FPL	42	4	No	Yes	Yes	Yes
16	MPL	FSG	FPL	10	1	No	No	No	Yes
17	MSG	FSG	MSG	0	0				
18	MPL	FSG	MSG	1	0				

In terms of animacy of the coordinated nouns, there seems to be some indication that a sentence with a coordination of singular animate nouns is infelicitous if a feminine plural adjective adopts a strategy of CCA (e.g. sentence 18 but not sentences 6 and 19):

(18) *O professor e a aluna escolhidas ...
The.MSG teacher.MSG and the.FSG student.FSG selected.FPL

(19) Tratamentos como a quimioterapia podem deixar o homem e a mulher estéreis
Treatments like the chemoteraphy can leave the.MSG man.MSG and the.FSG woman.FSG sterile.PL

The cases of feminine adjectives in CCA of gender with singular nouns found in the corpus were overwhelmingly of conjuncts involving inanimate nouns. For these cases although all the adjectives

used have an inherent gender, they are all compatible with inanimate nouns. Some of these adjectives are: *prematura* (premature.FSG), *irrestrita* (irrestrict.FSG), *típica* (tipical.FSG), *características* (characteristic.FPL) and *novas* (new.FPL). The same bias is not found when plural animate nouns are coordinated, where cases can be found of a feminine adjective in CCA with them. Whether for singular nouns this pattern is exclusively allowed for inanimates remains to be confirmed. However, unlike languages like Ndebele Moosally (1999) and Roumanian Farkas and Zec (1995), which have differential resolution according to whether the coordinated nouns are animate or inanimate, other agreement strategies are possible for inanimate nouns in Portuguese.

The results obtained also indicate that cases of CCA are not limited to synonyms or enumerations, but apply in other circumstances too:

(20) ... a percepção que toda sociedade faz sobre o plano e a
... the perception that the whole society makes about the.MSG plan.MSG and the.FSG
realidade econômica
reality.FSG economic.FSG
... the perception that the whole society makes about the economic plan and the reality

To summarise, the corpus data gathered suggests that a strategy of CCA can be frequently found in Brazilian Portuguese, and at least for this corpus, as frequently as a strategy of resolution. CCA is not restricted to occur with enumerations or synonyms, but can be found with other cases as well. However, although CCA can be widely applied, it does not seem to be applicable to coordinations of singular animate nouns, making such sentences infelicitous.

One obstacle faced in this investigation is that although coordinations can be frequently found in corpora, those involving nouns with different genders, and with postnominal adjectives scoping over both conjuncts are much less numerous, as can be seen by the difference between the initial and final frequencies in these tables. This is the case even in a corpus as big as NILC, where only around 15% of the sentences fulfilled these two constraints. The limited availability of annotated data means that for some of the issues under investigation there was not enough data for providing conclusive evidence. For instance, although the analysed data provided evidence for the correlation between animacy and CCA in the coordination of singular nouns, discussed above, for some cases the sample analysed is not large enough (e.g. cases 3 and 16 in both tables) for the hypothesis to be confirmed. For these cases it may be necessary to turn to the largest (albeit unannotated) corpus available for NLP, the World Wide Web. For instance, for case 16, CCA of gender and resolution of number, even though the analysed data contained only 1 sentence, a preliminary investigation using the WWW provided more positive evidence, as reported in Villavicencio and Sadler (2005).

In the next section we discuss the implications of this corpus study for a theory of agreement.

4 Capturing Agreement Patterns

Closest conjunct agreement has been discussed by Corbett (1991), Sadler (1999), Moosally (1999), Abeillé (2004) and Yatabe (2004) *inter alia*, and it is a strategy of partial agreement that can be found in many languages such as Ndebele (Moosally, 1999) and Welsh (Sadler, 1999). Moosally (1999), for instance, proposes an HPSG formalisation for capturing partial agreement in Ndebele, where agreement constraints are defined in a multiple inheritance hierarchy capturing agreement with the last conjunct, while Yatabe (2004) formalises CCA as part of a unified treatment which also deals with

coordination of unlike categories. However, the Portuguese data discussed in the previous section indicates that, in order to capture cases like that in (21), where mixed gender nouns are coordinated, and they trigger masculine agreement with the determiner and feminine with the postnominal adjective, it is essential to take into account information about the conjuncts in both extremities.

(21) Esta canção anima os corações e mentes brasileiras.
This song animate the.MPL heart.MPL and mind.FPL Brazilian.FPL
'This song animates Brazilian hearts and minds.'

Assuming an HPSG formalisation such as that of Pollard and Sag (1994), the attribute CONCORD, which is closely related to the noun's inflected form, reflects the resolved gender and number of the coordinate structure. The value of CONCORD can be computed by adopting a resolution approach such as that of Dalrymple and Kaplan (2000), whereby if there is at least a masculine noun in the coordinate structure, CONCORD.GENDER is masculine. To account for cases of CCA it is important to store agreement information about the leftmost and rightmost noun conjuncts, introducing two additional agreement attributes: LAGR, for the leftmost conjunct, and RAGR for the rightmost conjunct. For a coordinate structure, the values of LAGR, RAGR may differ, since the first two reflect the agreement values of each of the edge conjuncts, and determiners and prenominal adjectives agree with the coordinate structure via LAGR, while postnominal adjectives agree via RAGR.

All of the agreement patterns discussed in section 3 must be taken into account when a theory of agreement is proposed. For a sentence like 21, both LAGR and CONCORD are masculine and RAGR is feminine and the correct agreement values are observed, since the adjective can either agree with RAGR or CONCORD, but it will correctly rule out sentence (22) as ungrammatical.

(22) *Esta canção anima as mentes e corações brasileiras.
This song animate the.FPL mind.FPL and heart.MPL Brazilian.FPL

This formalisation can also capture sentences like 11 and 14, which have CCA for gender, but resolved number agreement for the postnominal adjective, if GENDER agrees with RAGR and NUMBER with CONCORD. Sentences like 8 to 10 agree exclusively with RAGR, with CCA of gender and number, while others like 6 agree only with CONCORD, with resolution of gender and number.

CCA of number and resolution of gender is not possible for Portuguese (sentence 18). Indeed, this strategy is not discussed in the literature or found in corpora, but it should be ruled out by a theory of agreement.

Therefore, to capture the agreement patterns found in Brazilian Portuguese, a theory of agreement must only allow (a) CCA of gender and number (RAGR:GENDER and RAGR:NUMBER), (b) resolution of gender and number (CONCORD:GENDER and CONCORD:NUMBER) and possibly (c) CCA of gender and resolution of number (RAGR:GENDER and CONCORD:NUMBER), but not (d) resolution of gender and CCA of number (CONCORD:GENDER and RAGR:NUMBER).

5 Conclusions

In this paper we investigated agreement patterns found in Portuguese NP (and N) coordinations, with evidence for different strategies gathered from syntactically annotated corpora. We concentrated on gender and number agreement between nouns and postnominal adjectives. The results obtained show

that a complex picture of agreement patterns in Portuguese NPs emerges from the corpus study, which should be taken into account when proposing a theory of agreement. The use of corpora provided not only concrete information about the frequency of use of each strategy, but also gave the basis for determining the contexts in which they can be used. Based on these results we discussed some of the characteristics that a theory should have in order to capture the investigated data, storing information about the leftmost and the rightmost conjuncts.

As future work we intend to compare the occurrence of these strategies in Brazilian and Portuguese corpora, to see if they are restricted to Brazilian Portuguese or not, and if not, if there are any differences in the frequency with which each of these strategies occur. This investigation aims at formalising a general and crosslinguistic theory of agreement.

References

Abeillé, A., 2004. A lexicalist and construction-based approach to coordinations. In: Müller, S. (Ed.), Proceedings of the HPSG04 Conference. CSLI Publications, Katholieke Universiteit Leuven.

Camacho, J., 2003. The Structure of Coordination: Conjunction and Agreement Phenomena in Spanish and Other Languages. Kluwer Academic Publishers, Dordrecht.

Corbett, G. G., 1991. Gender. Cambridge University Press, Cambridge, UK.

Dalrymple, M., Kaplan, R. M., 2000. Feature indeterminacy and feature resolution. Language 76 (4), 759–798.

de Almeida Torres, A., 1981. Moderna gramática expositiva da Língua Portuguesa. Martins Fontes, Sao Paulo.

Farkas, D. F., Zec, D., 1995. Agreement and pronominal reference. In: Cinque, G., Giusti, G. (Eds.), Advances in Roumanian Linguistics.

Kathol, A., 1999. Agreement and the syntax-morphology interface in HPSG. In: Levine, R., Green, G. (Eds.), Studies in Current Phrase Structure Grammar. Cambridge University Press, pp. 223–274.

King, T. H., Dalrymple, M., 2004. Determiner agreement and noun conjunction. Journal of Linguistics 40 (1), 69–104.

Moosally, M. J., 1999. Subject and object coordination in Ndebele: and HPSG analysis. In: Bird, S., Carnie, A., Haugen, J. D., Norquest, P. (Eds.), Proceedings of the WCCFL 18 Conference. Cascadilla Press.

Pollard, C., Sag, I. A., 1994. Head-Driven Phrase Structure Grammar. The University of Chicago Press, Chicago, IL.

Sadler, L., 1999. Non-distributive features and coordination in Welsh. In: Butt, M., King, T. H. (Eds.), On-line Proceedings of the LFG99 Conference.

Villavicencio, A., Sadler, L., 2005. An HPSG account of closest conjunct agreement in NP coordination in Portuguese. In: Proceedings of the HPSG05 Conference.

Wechsler, S., Zlatić, L., 2003. The Many Faces of Agreement. CSLI Publications, Stanford, CA.

Yatabe, S., 2004. A comprehensive theory of coordination of unlikes. In: Müller, S. (Ed.), Proceedings of the HPSG04 Conference. CSLI Publications, Katholieke Universiteit Leuven, pp. 335–355.