Abstract

The paper presents theoretical aspects and practical issues related to the development of a grammatical relation’s system for corpus annotation. The grammatical relations are arranged on a default inheritance hierarchy based on syntactic and semantic features. Preliminary tests on the annotation of an Italian treebank (the Turin University Treebank) show that the system implements a reasonable trade-off between richness of the representation and tractability of the annotation task.

1 Introduction

Statistical methods for linguistic studies are supported by the increasing availability of machine-readable resources. The fact that these methods require very large volumes of data has elicited a huge effort in collecting corpora with a variety of forms of annotation: part-of-speech, syntactic and word sense tagging. The corpora of syntactically analysed sentences are known as treebanks.

A treebank annotation schema can be based on the explicit representation of different kinds of information, i.e. the grammatical relations.

Grammatical relations (aka grammatical functions or thematic roles) encode the associations between the semantic predicate argument structures and their surface constituent structures (Bresnan and Kaplan, 1982) playing a relevant role in the semantic composition of the sentence. Grammatical relations are universals, but they are encoded in each language in different ways, according to its morphological and structural features.

Major differences in the encoding of grammatical relations can be found comparing configurational and nonconfigurational languages. Configurational languages are fixed word order languages, such as English; nonconfigurational languages are free word order languages, such as Czech. In the former, the grammatical relations can be mostly identified on the basis of word order and phrase structures. In the latter, they can be identified from different syntactic markers, such as case and other inflectional features (Bresnan, 1982).

In the description of natural languages also, more or less prominence can be given to the representation of grammatical relations. A constituency-based representation groups words in larger and larger units (phrases) and does not explicitly represents relations. Instead, a dependency-based paradigm (Hudson, 1990) mainly relies on grammatical relations between words, yielding constituency as a side effect. Annotation schemata based on grammatical relations are considered more adequate for nonconfigurational languages, while constituency based representations are considered optimal for the configurational ones (cf. Skut et al., 1997). Variations on constituency-based and dependency-based paradigm are in use in three well-known treebank projects in the literature: the Penn Treebank (Marcus et al. 1993, 1994), the NEGRA Treebank (Brants et al. 1997, Skut et al. 1997), the Prague Dependency Treebank (Hajčková, 2000), developed for English, German and Czech respectively.
The presence of nonconfigurationality in all languages to some extent, and the theoretical and applicative relevance\(^1\) of functional and semantic information, has triggered the annotation of grammatical relations in both dependency-based and constituency-based treebanks. In order to make the corpus really useful for users, be they linguists or NLP systems developers, we have to design annotation schemata that allow us to add as much linguistic information as possible.

Nevertheless, to ensure the tractability of the annotation task, the number of grammatical relations used in existing treebanks is quite small. The Prague Dependency Treebank uses 25 functions (at the analytical level) and about 40 semantic functors (at the tectogrammatical level)\(^2\), the NEGRA Treebank uses around 40 functions, and the Penn Treebank uses less than 20 semantic roles as affixes of phrase tags\(^3\).

We have developed a dependency-based treebank schema, centred upon the notion of predicate-argument structure and giving a peculiar prominence to the representation of grammatical relations. The schema allows for a richer and more detailed annotation because includes a very large number of relations specialised on the basis of two major criteria (morphosyntactic and semantic) and organised in a hierarchical structure. This organisation can be seen as an *underspecification* mechanism because it provides relations with variable degrees of specificity. The usage of this mechanism in the practice of the annotation, on the one hand, ensures the tractability of the annotation task, also managing ambiguity and vagueness; on the other hand, it solves inter-annotators disagreement problems.

\(^1\) The theoretical relevance of the representation of grammatical relations has been pointed out in Lexical Functional Grammar (Bresnan and Kaplan, 1982), in Fillmore’s case grammar, in Perlmutter’s relational grammar, in Hudson’s Word Grammar (McCauley, 1988). Moreover the representation of grammatical relations can be very useful in a number of applicative tasks; such as Information Extraction (Vilain, 1999).

\(^2\) The schema of the Prague Dependency Treebank consists of three levels: morphological, analytical (surface syntactic structure) and tectogrammatical (deep syntactic structure revealing the topic-focus articulation with syntactic functors and attributes describing the contribution of each word in the communication act).

\(^3\) The annotation of the Penn Treebank has been augmented with a semantic layer described in Palmer et al., 2000.

This paper describes the theoretical aspects of the hierarchy and practical issues related to the development of the grammatical relation’s system of the Turin University Treebank (TUT), which has been empirically tested in the annotation of a corpus of non-restricted Italian texts. The next section presents the hierarchical organisation and the specialisation criteria of grammatical relations. Section three presents a more detailed description of the application of this system and describes solutions adopted to ensure the tractability of the annotation task and inter-annotator agreement.

### 2 Building a grammatical relation’s system

TUT adopts a dependency-based formalism taking relations between words as basic primitives (see Lombardo and Lesmo, 2000). Dependencies are directed grammatical relations linking pairs of words and the set of relations involved in a sentence, forms a *dependency tree*. The choice of this formalism is motivated by the advantages coming from the explicit representation of grammatical relations and predicate-argument structures, and since Italian is a partially configurational language. As quantitatively confirmed in a study conducted on a subset of the TUT corpus, in Italian declarative sentences all the six permutations of Subject Verb Complement are allowed\(^4\).

Starting from some theoretical issues, we have implemented the dependency relations and we have tested them during the annotation of an Italian corpus of non-restricted texts.

#### 2.1 The hierarchical organisation

A hierarchical structure offers a conceptual framework for the representation of information. Taxonomic hierarchies, well known in AI as *inheritance hierarchies*, organize information at appropriate levels by inclusion relations. The usage of inheritance hierarchies in NLP comes from three separate traditions: semantic networks in AI, object-orientation in computer science and the notion of “markedness” in linguistics. It is well motivated by the possibility

\(^4\) A preliminary study on around 400 sentences (10.000 words, 460 different verbs) shows that the more common order is S-V-C (68,3%), followed by S-C-V (12,4%), C-V-S (7,4%), V-C-S (6,4%), C-S-V (3,1), V-S-C (2,4).
of capturing linguistically interesting abstractions, representational compactness, ease of maintenance, uniformity of treatment of several conceptual levels, modularity and reusability (Daemelans et al., 1992). The concept of default inheritance is explicitly incorporated in several linguistic frameworks to model different layers of analysis: the lexicon (i.e. in Categorial Grammar (van der Linden, 1992)), the syntactic features in GPSG (Gazdar et al., 1985), all the layers in Word Grammar (Hudson, 1990; Fraser and Hudson, 1992).

We apply to our grammatical relations, defined as collections of properties, a hierarchical organisation, explicitly indicating that:

- all the relations inherit the properties of the most generic parent relation, DEPENDENT, such as the nature of grammatical relation representing some surface syntactic dependency;
- each relation (except the root) is a specialisation of its parent relation and is included in its subset (set of relations sharing all the properties of the parent);
- each relation (except leaves) is a generalisation of its children.

Each relation is formally defined showing its internal structure composed by features signed by unary predicates ‘+’ (true) and ‘-’ (false). A definition such as <+Complement, +Verbal-Dependent> is a well-formed definition of a relation indicating that the relation VERBAL-DEPENDENT is the parent of COMPLEMENT. Complement and Verbal-Dependent are D-features, the basic features of our system (see Fig.1), but a relation can be further specialised using other particular features.

2.2 Specialisation of grammatical relations

In our system, the further specialisation of relations is driven by two major criteria:

- **morphosyntactic criterion** (M-criterion): the morphological category of one of the words involved in the relation determines its Morphological-extension (M-extension);
- **semantic criterion** (S-criterion): the Semantic-extension (S-extension) of a relation depends on some semantic feature of one of the words involved in the relation.

The *M-extension* of a relation R is obtained adding an M-feature to the features of R. R is *M-extended* if there is an M-feature in its representation, or there is an M-feature in the representation of a parent from which R inherits. R is *M-extendable* only if it is not M-extended. The definitions of *S-extension* and related concepts are analogous. This means that each relation can be M-extended or S-extended only once, i.e. there are no more than one feature of morphosyntactic and semantic type in each relation representation. The sets of basic features, M-features and S-features are disjoint: syntactic features are semantic primitives, such as Time, Location, and Age.

2.2.1 The morphosyntactic criterion

The first criterion reflects the theoretical classification of linguistic non-relational concepts made in Hudson, 1990, where the basic classification of words is based on the word-type (or grammatical category, such as noun, verb, adjective, ...). The word-type is a non-relational concept, whereas the other features of words are relational concepts (Hudson 1990; Bresnan, 1982).

In the practice of annotation, the usefulness of this criterion consists in allowing for an explicit representation of different behaviour of words belonging to different categories. Using the M-criterion, we define, for example, relations such as ADJCMOD (adjectival modifier) (i.e. in ‘interesting argument’) or ADVBMOD (adverbial modifier) (i.e. in ‘more interesting’). M-extensions of MODIFIER represented as <+Adjcmod, +Modifier, +$Adjective> and <+Advbmod, +Modifier, +$Adverbial>.

2.2.2 The semantic criterion

The second criterion reflects the distinction between generic grammatical functions and semantic functions (Bresnan, 1982).

There is not a unique universally accepted set of

---

6 A major basic distinction in Word grammar (Hudson 1990) is stated between relational and non-relational categories. WG syntax is centered on two inheritance hierarchies, one for word types and other for grammatical relations ... the category ‘word’ is basic in every sense (Fraser and Hudson, 1992).

5 All the features not specified can be seen as false.
Figure 1. The hierarchical organisation of grammatical relations.

Semantic primitives, and we have adopted a set of around one hundred semantic suffixes including two main kinds of semantic functions:
- very specific semantic functions, which can be easily identified, such as AGE, or TRANSPMEANS (to indicate the transportation means used for travelling);
- traditional semantic functions, with semantic specifications that are well represented in the literature, such as LOC (for location) or THEME.

When the S-criterion is applied to verbal dependents, it is useful in the identification of semantic roles. In general, it seems desirable to label each argument of a predicate with an appropriate semantic label in order to identify how sub-constituents are semantically related to their predicates (identification of verb subcategorization frames). For instance, we can define AGTCOMPL (agent complement) (i.e. in ‘symphonies recorded by Toscanini’), which S-extends COMPLEMENT, <+Agtcompl, +Complement, +Agent>.

If the S-criterion is applied to non-verbal dependents it allows also for the representation of other semantic information relevant in syntactic representation. For instance, we can define PREPMOD-AUTHOR (prepositional modifier author) (i.e. in ‘a book of Grisham’) or PREPAJT-TRANSPMEANS (prepositional adjunct transportmeans) (i.e. in ‘…arrived by plane’). PREPMOD-AUTHOR, <+Prepmod-Author, +Modifier, +Preposition, +Agent>, is both M- and S-extension of PREPMOD; PREPAJT-TRANSPMEANS <+Prepajt-Transpmeans, +Adjunct, +Preposition, +Transpmeans> is M- and S-extension of ADJUNCT.

To increase the readability of annotation, the names of relations are built according to the features that define and extend it. Exceptions of these naming rules are relations, who are present with a traditional name in the literature, for example, the argument of a determiner (DETARG), usually a common noun, is named NBAR; verbal heads can have one, two or three complements that have well-established names in the literature: subject (SUBJ), direct object (OBJ), indirect object (INDOBJ), etc.

2.3 Higher levels of the hierarchy

Starting from the most generic idea of grammatical relation, DEPENDENT, we draw some fundamental distinctions in D-features. A first important distinction keeps apart coordination (COORDINATION), punctuation markers (SEPARATOR), a particular set of relations we call EXTRA (that collects relations of various natures), verbal (VERBAL-DEPENDENT) and non-verbal dependents (NON-VERBAL-DEPENDENT).

COORDINATION generalises all kinds of relations that can be involved in coordinative structures, i.e. COORD, which links a Conjunction to its head (first conjunct), COORD-2ND, which links the second conjunct to the Conjunction. SEPARATOR is the most generic relation used in punctuation marking. EXTRA generalises all those relations that cannot be easily classified under the other part of the hierarchy because of their atypical behaviour. It includes very different relations such as APPPOSITION, VISITOR.

---

7 The representation in dependency-based formalism of coordinative structures is particularly problematic and different approaches are reported (see Hudson 1990, Mel’cuk 1986). Our approach (Lombardo and Lesmo, 1998), privileging one of the two conjuncts as a head of the whole coordination, is motivated by the non-reversibility of coordinative structure in which the syntactic differences between the two conjuncts are taken into account (see Mel’cuk 1986).

8 This is the relation between an extracted word and the verb from which it depends. The idea of recognizing an explicit relation between the extracted and the first verb is more familiar in constituency-based theories than in dependency based ones. (Hudson, 1990)
All other relations appearing in the hierarchy are specialisation of VERBAL or NON-VERBAL-DEPENDENT. In both VERBAL- and NON-VERBAL-DEPENDENT relations, we have introduced the distinction, well known in the literature, between complements and adjuncts. A complement is obligatory and closely linked to its head, an adjunct is optional and only loosely linked to its head. Moreover the head itself determines the semantic relation between the head and its complement (subcategorises the dependent); whereas the semantic relation between the adjunct and its head is determined by the adjunct (Hudson, 1990). In our taxonomy, we call ARGUMENT a non-verbal complement and MODIFIER a non-verbal adjunct; we call COMPLEMENT and ADJUNCT the verbal dependents (see Fig.1).

3 Application of the grammatical relation’s system

The usage of the specification mechanism based on features, gives peculiar richness to our grammatical relation’s system, which can specify a large number of different relations. Nevertheless, the problems that arise in the annotation of relations are worsened using a richer schema, because the selection of the correct grammatical relation can be more difficult navigating in a search space consisting of a large number of competing labels. Moreover the specificity of relations can increase also the inter-annotator disagreement. These problems are approached according to the hierarchical and flexible organisation of relations which allows a fine solution: when an annotator is uncertain among multiple solutions for a dependency label, the solution is to climb up the hierarchy and to assign a higher label, at some level where the annotator feels confident. We can deal with inter-annotators disagreement problems (i.e. when two annotators label a syntactic dependency using two different relations) in an analogous way, finding the common most specific ancestor of two relations. By separating D-, M- and S-features, and postulating that only one semantic feature and one syntactic feature can be associated to each relation, we adopt the first strategy (as in WG). Allowing for the underspecification of relations and organising them in a multiple default inheritance system we ensure a trade-off between accuracy of description and tractability of annotation also giving a solution to the inter-annotators agreement problem. The underspecification of the M-feature can be useful, for instance, in the annotation of constructions where are present syntactic hybrid of two different category types, such as Italian infinitive noun (*infinito sostantivato*). The

---

9 Like in GPSG (Gazdar et al., 1985), a feature-based theory where a syntactic category can be accepted also if some of its features are not specified.

10 The problem of head sharing constructions and the case of *infinito sostantivato* is reported in (Bresnan, 1997).
underspecification of S-feature can be instead useful in semantically ambiguous constructions.

4 Conclusions

In this paper we present and motivate the grammatical relation’s system developed for the annotation of the TUT. Richness and flexibility of annotation are the major peculiarities of this system. Problems related to inter-annotators agreement and specificity of annotation are approached by means of a careful hierarchical arrangement of grammatical relations. Preliminary tests have been performed on a corpus annotated using this system with programs for the extraction of subcategorization frames. Future applicative development of the TUT project will give empirical validity to the approach here described.

References


