Using feature structures as a unifying representation format for corpora exploration

Julien Nioche
LexiQuest & Université Paris X Nanterre
julien.nioche@lexiquest.fr

Benoît Habert
LIMSI-CNRS & Université Paris X Nanterre
habert@limsi.fr

Abstract

In this paper we report on the use of feature structures to represent the linguistic information of a corpus. This approach has been adopted in TyPTex, a project which aims at providing a generic architecture for corpora profiling. After a brief overview of the Typtex project, we show that corpora exploration requires manipulating linguistic features in order to obtain a required level of linguistic information or changing the set of features to get a new point of view on the data. We show that feature structures formalism can help the building and management of linguistic features with Meta-Rules based on unification. Finally, we provide an example of marking which uses a mixed approach between projection of information from a static lexicon and contextual marking via Meta-Rules. Results tend to show that the use of feature structures can improve the coverage and reliability of the marking.

1 Introduction

Huge tagged or parsed corpora are used in a broad number of language-related studies (McEnery and Wilson, 1996), with very different goals, such as lexical acquisition, speech processing, language learning or discourse analysis. The common point of these works is the annotation of these corpora according to linguistic features. Those features may be of various kind, going from simple morphological or lexical information to more complex grammatical, semantic, functional or phonetic features. Global information (like the number of words, the average word length, etc.) can be also used for corpora processing. Some kinds of features can be easily identified in texts, such as morphological information which can be obtained with simple *n-gram* computing, or lexical features which requires only word segmentation. On the other hand, more complex tools or resources may be necessary to get the awaited level of information. This is the case of syntactic analysis, obtained with a parser, and of semantic features, which require dictionaries, or anaphora resolution.

Depending on the goal of the study, the set of features observed in corpora changes. For example, readability measures consider the length of words and their average frequency compared to frequencies from a reference corpus, while stylometry studies, the stylistic analysis of texts for the purpose of author attribution, require other kind of linguistic information, like lexical features, syntactic constructions or textual organization. Therefore the choice of features is task dependent. Some corpus-based studies need to combine different kind of features at the same time. This is typically the case of distinguishing among language registers (Biber 1993, 1995) and style (Tambouratzis et al., 2000). For example, Tambouratzis et al. (2000) combine morphological, lexical, grammatical and structural features to bring out style differences within a Greek corpus. Sets of heterogeneous features are thus put together in order to compare subparts of a text collection. This kind of studies concerns *corpora exploration*, since the choice of a feature set changes the results obtained and reveals different aspects of the data. Thus linguistic features must be manipulated easily, which raises the issue of their representation.

A lot of formalisms are used to handle linguistic annotation[1]. They are nowadays mostly based on manipulation of XML/SGML entities. Tools are also provided in order to create, search or browse corpora within these formalisms. The lack of standards to represent and manipulate linguistic information is a problem for Natural Language Processing, since processing corpora out of their foreseen use or in combination with textual resources in a different format requires an extensive work to build conversion and/or manipulation tools. Attempts are made to solve this problem. The AMALGAM (Atwell et al., 1994) project aimed at developing methods of automatically mapping between the annotation schemes of the most widely known corpora, for both POS tag sets and phrase structure grammar schemes, to improve the reusability of the data. More recently, Bird et al. (2001) propose a formal framework for linguistic annotation, in the context of Speech Processing. This

[1] See the Linguistic Annotation Page at *www.ldc.upenn.edu./annotation.*

framework aims at providing a core representation format, which regroups the common features of existing annotation schemes. In the domain of morphosyntactic annotation, the EAGLES project recommends a common formalism[2], where information is represented by the position of characters in a tag. There is an obligatory position for the POS features, which has a closed set of possible values. Other linguistic features can be freely encoded with a symbol within the tag.

This paper explores the benefits that feature structures offer for the representation and manipulation of linguistic information in corpora. Experience on corpora profiling in the TyPTex project shows that the use of this format helps to handle information in a clean way, to manipulate and modify sets of linguistic features, and improves reusability of both data and experiences. Using feature structures can also improve the marking of linguistic phenomena.

2 Overview of the TyPTex project

The goal of the TyPTex project is to provide a generic architecture for corpora profiling. This project is financed by ELRA (*European Language Resources Association*) and is carried out jointly at LIMSI and UMR 8503[3]. Work within this project has been previously described by Illouz et al. (2000) and Folch et al. (2000).

2.1 Background

The underlying idea is that the reliability of the knowledge acquired form a corpus depends on the homogeneity of its data and is decreased by its heterogeneity. In the domain of morphosyntactic tagging, Biber (1993: 223) used the LOB (*Lancaster-Oslo-Bergen*) corpus to show that the probability of occurrence of a morphosyntactic category depends on the domain of the text. The same is true with sequences of morphosyntactic categories, which frequencies vary according to the domain. Sekine (1997) compared the performances of a probabilistic syntactic parser with different configurations for training and testing, using 8 domains of the Brown corpus. This work proved that the quality of the parsing in terms of precision and recall falls as the domains of the texts used for training and testing differs. Ruch and Gaudinat (2000) compared the lexical ambiguity between medical and general texts and underlined the necessity to build domain-adaptable tools for Natural Language Processing. These studies lead to the conclusion that the use of important corpora requires profiling tools in order to get indications about lexical and morphosyntactic uses of their subparts and thus determine their homogeneity or heterogeneity. Corpora profiling and tuning can globally improve the performances of NLP tools, as shown in Illouz, (2000).

2.2 Previous works

The approach in TyPTex consists in developing a typology of texts through inductive methods. It means that the text types are defined in terms of sets of correlated linguistic features obtained through multivariate statistical techniques from annotated corpora. This approach is based on Biber's (1988, 1995) work. Biber uses 67 features corresponding to 16 different categories (verb tense and aspect markers, interrogatives, passives, etc.). He examines their distribution in the first 1.000 words of 4.814 contemporary English texts from reference corpora. The identification of the 67 features in the corpus is done automatically on the basis of a preliminary morphosyntactic tagging. The accuracy of the tagging is checked by a linguist. The sets of correlated features (the dimensions) are obtained through a multivariate statistical technique (factor analysis). Each dimension consists of two complementary groups of features which can be interpreted as positive and negative poles. In other words, when one group of features occurs in a text, the other group is avoided. Statistical methods are then used to group texts into clusters according to their use of the dimensions. These clusters correspond directly neither to text "genres" nor to language style or registers.

2.3 Data, tools and methods

The corpus used in TyPTex to test and tune the system represents 5 million words and is a part of the corpus gathered by G. Vignaud (INALF – Institut National de la Langue Française) and B. Habert within the European project PAROLE[4]. The texts are tagged according to the *TEI* (*Text Encoding Initiative*) recommendations. Queries are then performed to extract a subset of texts which are relevant for a determined study or application. The next step is to achieve a morphosyntactic tagging which

[2] Available at *http://www.ilc.pi.cnr.it/EAGLES/annotate/annotate.html*.
[3] See *http://www.limsi.fr and http://www.ens-lsh.fr*.
[4] See *http://www.elda.fr/Fr/cata/doc/parole.html*.

associates each lexical item (or polylexical item) with a given lemma, a part of speech and other morphosyntactic information. The tagger used currently is Sylex-Base. It is based on the work of P. Constant (Ingenia, 1995), and proved to be robust during the tagger evaluation program GRACE (Adda et al., 1998).

The second step is *typological marking*. It consists of replacing the information generated by the morphosyntactic tagger with higher-level linguistic features. These new features are obtained on top of the morphosyntactic tags and vary according to the oppositions the user wishes to bring out. Section 3 will explain more in detail why and how such manipulations of linguistic features are effected. From the resulting marked corpus several matrices are generated, in particular the matrix containing the frequencies of each feature in each text of the corpus under study. The resulting matrix is then analysed by statistical software programs. The analysis of the matrix aims at, on the one hand, identifying features that reveal a certain kind of opposition among the subparts of the corpus, and on the other hand, making an inductive classification of texts.

3 Corpora exploration and manipulation of linguistic features

A lower level tagging used in TyPTex includes shifters, modals, presentatives ("*il y a"* and "*c'est*"), tense use, passives, certain classes of adverbs (negation, grading), determiners, etc. From the features tagged initially (around 300 available with Sylex and 170 with Cordial (Synapse, 1998)), about 40 were kept and divided into 2 subsets. The first subset comprises functional elements which role is the organization of discourse and sentence. The second subset comprises open categories like nouns, adjectives or verb tense.

The features available with the initial POS tagging may not be sufficient for a given study. There is often a gap between what one gets at the output of a tagger and what is aimed at. In TyPTex, we call typological marking the set of features that is presumed useful to bring out different types of texts. Features has to be manipulated in order to get this awaited level of typological marking. However a set of linguistic features can not be settled once and for all. Typological marking requires a lot of explorations : one needs to test a set of linguistic features by analysing the distinctions it brings within a corpus.

Sometimes features can be too fine-grained and lead to a scattering of occurrences which makes contrasts imperceptible. This was the case in one of the pilot studies (Illouz et al., 1999) for the TyPTex project with the verb category, which was divided into some 50 features (due to the morphology of French). Most of those features had a only a few occurrences and were not statistically significant. The other problem with this splitting of the features was that it offered no indication about the use of the verb in general. Thus it was impossible to check whether a under-use of Nouns in a subpart of the corpus was related to an over-use of Verbs in the same subpart. This is why some elementary features had to be regrouped inside "super features", covering larger categories which were not available with the initial tagging.

Some features can also be too rough and hide real oppositions. For example, the same tag can be used by a tagger to cover indifferently quantity indicators, as well as dates. We can presume that splitting that general "*Cardinal Number*" feature into two sub-features (quantity and dates) would create finer distinctions among the corpus. For instance, it can be necessary to gather tags corresponding to the same function but belonging to different morphosyntactic categories, such as some punctuation marks and some conjunctions, which can be regrouped as textual markers. For example, this could be the case of punctuation symbols marking an incision in the discourse, such as quotes, parenthesis or long dashes.

A good illustration of this point is the use of the distinction between qualitative and relational adjectives made by Habert and Salem (1995). The aim of their study was to reveal differences of language use in a sociologic corpus of open answers. A first feature set, using "traditional" POS features (Verb, Adjective, Noun, etc.), revealed a difference in the use of verbs (overused by less-educated persons) and nouns (overused by educated persons) between two different groups of answers. At this stage, *adjective* was considered as an "atomic" feature. A modified set of features introduced an opposition between *qualitative* and *relational* uses inside the *adjective* category. Relational adjectives are sometimes called "pertainyms", since they mean something like "of, relating/pertaining to, or associated with" some noun and play a role similar to that of a modifying noun (Fellbaum, 1990). For example, *geographical* in "*a geographical map*" refers to *geography*, as *presidential* in "*presidential election"* is linked to *president*. Adjectives that are not relational are considered to have a "qualitative" function (which modifies the quality of a noun), such as "*nice"* in "*a nice child*" or "*good"* in "*a good*

practice". This notion of relational use of adjectives is used in WordNet[5] (Miller, 1998), where relational adjectives are linked to their corresponding name, the other adjectives being mainly studied through antonymy and contrast. This distinction enabled to refine the cluster of "educated persons" between non-graduate and graduate persons. In that case, splitting the *adjective* category into two finer categories provided a better description of the corpus.

Feature manipulation is required to get the awaited level of linguistic information, by completing and modifying the feature set available with the initial POS tagging or changing the features retained for typological marking. It is necessary therefore to be able to group features for one contrast, to divide others, and at times even to start afresh tagging and marking. Corpus exploration requires flexibility in the manipulation of the feature set which in turn introduces constraints on the representation formalism. In the Typtex project, feature structures are used to represent linguistic features of the words in the corpora.

4 Using feature structures as representation format

4.1 From tags to features

A morphosyntactic tagger associates a given piece of information with a lexical token by the way of a tag. According to the software used, the content of this information may vary, and its form as well.

action action Ncfs (CORDIAL)
"action" nom : féminin singulier (SYLEX)

Figure 1 Examples of different POS tagger outputs

Figure 1 presents the output of two taggers CORDIAL and SYLEX for the word "*action*", where the linguistic information is the same (except for the lemma, which is not present here with Sylex) but changes in form. In fact, there's often a confusion between the graphical forms that programs manipulate (tags) and the linguistic features used by the linguist. Tags may be different but represent the same linguistic feature. In the TyPTex project, an intermediate format is used to represent linguistic information, independently of the tags provided by the preceding process (POS tagging, for example).

4.2 Representation

Feature structures such as those employed in unification grammars are used in order to represent the linguistic information contained in a corpus. The format used in TyPTex is inspired from the PATR formalism (Shieber, 1986). A feature structure associates values with a set of features, and can be represented by an equation, where the feature is written between < > and the value is placed after a symbol *equals* (=). Feature structures can be *atomic* (one feature associated with a value) or *complex* (a value is itself a feature structure, in a recursive way). The example of tagging provided above will give the following feature structure :

<form> = action
<lemma> = action
<category> = noun
<type> = common
<agreement gender> = feminine
<agreement number> = singular
.

Figure 2 Equation of a feature structure

In this example, some features have an atomic value (<form> for instance), whereas another as a complex one (<agreement>). Feature structures can also have a graphical representation with a DAG (Directed Acyclic Graph), as on figure 3.

[5] See *http://www.cogsci.princeton.edu/~wn/*.

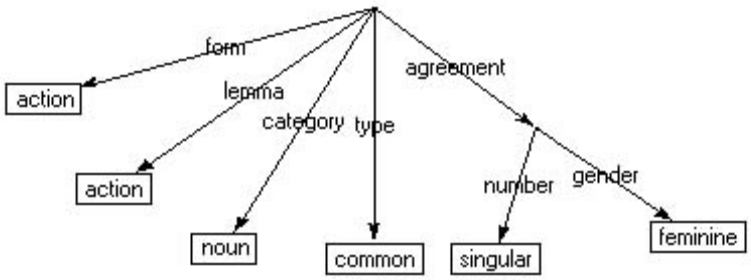

Figure 3 DAG representation of a feature structure

4.3 Modifiers

Modifying operators can also be used in the equation. This is the case of *negation*, marked by a tilde ~, which implies that a feature shall not have a given value, as in : <*agreement gender*> = ~*neutral*. *Disjunction*, is an other aspect. It is the possibility to associate multiple values with a feature, by putting these values between braces { }[6]. At last, two or more structures can share the same value (*co-indexation*) when it is placed between parenthesis. This enables, for example, to specify an agreement in *number* and *gender* between an adjective and a noun. In that case, the *agreement* features of the adjective and the noun are not only equals, but share the same value.

4.4 Use in the TyPTex project

Each word of the corpus used in TyPTex is represented by a feature structure. Thus, it is possible to modelise more precisely the information resulting from marking, in the style of Gazdar et al. (1988). This kind of representation format allows one to manipulate linguistic features directly instead of tags. A mapping (comparable to the mapping done in Atwell et al., 1994) is effected between the output of the morphosyntactic tagger used upward and the corresponding linguistic feature structures, the downward processing gaining in independence from the format of the tagger used. It is possible to use an other program for the POS tagging-task and keep the same level of information thanks to a mapping into feature structures. This enables to compare the results provided by different taggers and see their influence on the quality of the marking.

An other advantage is that feature structures can encode any type of linguistic features, at any level, morphological, syntactic, semantic or functional. Compared to a tag, a feature structure has the following qualities :

1. Linguistic information is named explicitly and does not depend on the position of a character inside a linear string, like in a tag. A feature structure (ex: <*category*>=*noun*, <*type*>=*common*, <*gender*>=*feminine*, <*number*>=*singular*) is easier to read than a tag and less ambiguous (ex: *Ncfs*).
2. Linguistic information is structured and hierarchized. A structure such as (<*category*>=*noun*, <*type*>=*common*, <*agreement gender*>=*feminine*, <*agreement number*>=*singular*), is more logical than (<*category*>=*noun*, <*type*>=*common*, <*gender*>=*feminine*, <*number*>=*singular*). In this example, the sub-features <*gender*> and <*number*> can be manipulated through the whole complex feature <*agreement*>.

However, one of the most remarkable aspects of the feature structure formalism is that it provides useful mechanisms for feature manipulation, which is strongly required for corpora exploration.

5 Manipulation of data with Meta-Rules based on unification

Since some linguistic features one wants to obtain for typological marking may not be available at the output of a morphosyntactic tagger, it becomes necessary to add, retrieve or modify some features to get the awaited level of information for typological marking. On top of that, corpora exploration requires to test several feature sets in order to bring out new contrasts inside the corpus, which also implies the use of tools for features manipulation.

[6] This is a good way to represent the output of a non-deterministic tagger, which can give different possible tags for a single token.

5.1 Unification of feature structures

Feature structure are based on the concept of *unification*, which can be defined by the following : *"Unification of two feature structures A and B is the minimal structure containing both A and B"*. Unification checks the compatibility between two feature structures (a feature must not have two different values) and when possible produces a structure containing all the information of both structures.

For example, the unification of the structure :

<category> = noun
<type> = common
<agreement gender> = feminine

with the structure

<category> = noun
<type> = common
< agreement number> = singular

yields:

<category> = noun
<type> = common
<agreement gender> = feminine
< agreement number> = singular

5.2 Meta-Rules

Meta-Rules (as in Jacquemin 1994a, 1994b and 1997) based on unification are used to modify feature structures. Basically, a Meta-Rule is a feature structure consisting in two parts : a source part and a target part (in an equation, the source is separated from the target by the symbol "=>"). A Meta-Rule is applied on the feature structures representing a word (or a sequence of words) of the corpus : if unification is possible between the source part of the rule and the feature structure of the corpus, then the last is replaced by the target part of this rule. Thus Meta-Rules do not only add information to a given feature structure of the corpus but totally rewrite it.

The role of unification is to check the compatibility between the source part of the Meta-Rule and a word of the corpus. However the unification process used here slightly differs from usual unification. It could be called *constrained unification*, because it stipulates that the feature structure representing the word must contain at least all the information of the source part of the Meta-Rule (same features with compatible values). Thus, the source part of the Meta-Rule *subsumes* the feature structures of the word. This condition (unification + subsumption) constraints the writing of the rules and ensures that they are used in the correct cases.

For example, using normal unification with the Meta-Rule :

<tense>=present
=>
<tense>=present
<deictic>=yes

changes any Noun represented minimally by the feature structure :

<category> = noun
<type> = common
< agreement number> = singular

into :

<tense>=present
<deictic>=yes

The use of constrained unification prevent such cases.

5.3 Positional information

One important aspect of using feature structures is that it can help to modify the linguistic information of a corpus in a contextual way. In this case word tokens are not submitted to Meta-Rules separately, but inside sequences (corresponding to a paragraph, for example). Thus, the scope of a Meta-Rule can be larger than one slot in the sequence and covers several words. One tries to apply each Meta-Rule starting from each position inside the sequence (by checking whether the source-part of the Meta-Rule subsumes and can be unified with a part of the sequence of feature structures). All Meta-Rules are tested on position 1 in the sequence, then on position 2, and so on until the end of the sequence. In case of success, the slots inside the sequence covered by the source part are replaced by the target-part of the Meta-Rule.

Positional information is added to the Meta-Rule in order to specify the distance between the words in the sequence. This information is present in the feature structures of the rule as a feature by itself. The notation depends on the type of the distance, which can be fixed (noted by a *"p" + number of the token in the sequence***,** ex: *p4*), free (noted by a "*" + *number of the token, ex: *4*), or limited (noted by a "*" + *number of the token* + "+"+ *maximal distance*, ex*: *4+3*). The position 1 corresponds to the current position of the processing inside the sequence.

```
<p1 lemma>="
<p1 category >=punctuation
<*2+5  lemma>="
<*2+5 category >=punctuation
=>
<p1 lemma>="
<p1 category >=punctuation
<p1 type>=start_short_citation
<*2+5 lemma>="
<*2+5 category >=punctuation
<*2+5 type>= end_short_citation
.
```

Figure 4 Distances in Meta-Rule

Figure 4 shows a Meta-Rule, which adds a feature <*type*> to the quotes of a text, with a value *start_short_citation* or *end_short_citation*, if there is a distance of up to 5 words between the positions of the quotes. In that example, the distance is limited to 5. This Meta-Rule could be used to differentiate cases of reported discourse or to distinguish citations marking a phenomenon of distanciation form the speaker.

5.4 Meta-rules in TyPTex

Meta-Rules are used in order to manipulate linguistic features by regrouping them into larger categories, or in the opposite splitting into finer features. This is a convenient tool for managing the feature sets used in a study and for comparing the results that they provide. It also helps to add features not available at the output of a tagger and thus get the level of information awaited for typological marking. One of the most powerful aspect of the Meta-Rules is that they are *contextual* : one can manipulate the content of a feature structure (basically a word) dependently of the context of that structure (other words in the same sentence or paragraph). In the rest of that paper, we will show that the use of Meta-Rules can improve the marking of such a subtle distinction as the one between qualitative and relational adjectives.

6 An application case: distinguishing between qualitative and relational adjectives in a corpus

In the Typtex project, an opposition is projected in the corpora between relational and qualitative adjectives (see section 3). The description of relational adjectives provided by Habert and Salem (1995) is followed :

1. they are equivalent to a sequence of nouns : *presidential election / election of the president*
2. they are never gradable : **a very geographical map*
3. they cannot have a predicative function : **response to the virus was immune*.

We realized that the addition of a distinction between qualitative and relational adjectives improved the description of a corpus in Habert and Salem (1995) by refining the groupings of texts obtained after a multivariate analysis. But although this opposition seems to be useful for corpora description, its use is far from being obvious. As noticed by Bartning and Noailly (1993), a lot of adjectives can be analysed either as relational or qualitative, depending on the context. This is the case with the French adjective *économique*, which has a relational function in "*la politique économique*" (related to economics), but a qualitative one in "*une formule économique*"(which is not expensive). One can even assume that any relational adjective can take a qualitative function in a given context. That is why the distinction between these two aspects is somehow difficult to process automatically without information about the context of use.

The solution adopted in Typtex was to combine two approaches to distinguish between qualitative and relational adjectives : on the one hand a list of potentially non-ambiguous adjectives was used to annotate the corpus (information is *projected* on the texts) while on the other hand a set of Meta-Rules was intended to disambiguate the adjectives in context (information is then *extracted* from the texts). Here we compare these two approaches.

6.1 An empirically-build static list

A list of relational and qualitative adjectives has been constituted manually using press articles of the French newspaper *Le Monde*, and taken from the PAROLE corpus. The 14 million words subpart *Press* of the corpus has been built by random selection of full issues of *Le Monde* and gathers issues from 1987, 1989, 1991, 1993 and 1995. To build the list we extracted the thousand most frequent adjectives of the whole corpus *Le Monde* and analysed them manually, to check whether they have a priori a relational or qualitative function out of context.

Relational	Qualitative	Ambiguous
annuel	absolu	commercial
automobile	actuel	français
bancaire	ambitieux	historique
budgétaire	ami	humanitaire
cardinal	ancien	idéologique
constitutionnel	beau	judiciaire
exécutif	bien	logique
…	…	…

Figure 5 A sample of the adjective list

In its final state the list contains 264 non-ambiguous adjectives with 244 qualitative and 20 relational. All the other adjectives studied were judged too dependent on the context and thus ambiguous. Figure 5 shows a sample of this list.

6.2 Description of the Meta-Rules

Afterwards we created a set of Meta-Rules using feature structures for disambiguation. These rules are the following :

An ambiguous adjective is considered as qualitative if :

1. it is directly preceded by a grading adverb
2. it is directly preceded by a stative verb
3. it is directly preceding a noun, with the same value of number and gender[7]
4. it is directly preceded by a qualitative adjective and a conjunction
5. it is directly preceding a conjunction and a qualitative adjective
6. it stands alone between two double quotes

An ambiguous adjective is considered as relational if :

7. it is directly preceding a conjunction and a relational adjective
8. it is directly preceded by a relational adjective and a conjunction

[7] This is at least true in French, where the only position of the relational adjective is after the noun.

Rules 4, 5, 7 and 8 are based on the hypothesis that coordination is only possible between adjectives sharing the same function as whether qualitative or relational (e.g., compare *young and nice* with **beautiful and geographical*). Hatzivassiloglou and Wiebe (2000: 300) report on similar property of conjunctions for assigning semantic orientation to adjectives (e.g., compare *corrupt and brutal* with * *corrupt but brutal*).

Figure 6 shows the equation corresponding to rule 1.

```
<p1 form>=(1)
<p1 lemma>=(2)
<p1 category>=adverb
<p1 type>=grading
<p2 form>=(3)
<p2 lemma>=(4)
<p2 category>=adjective
=>
<p1 form>=(1)
<p1 lemma>=(2)
<p1 category>=adverb
<p1 type>=grading
<p2 form>=(3)
<p2 lemma>=(4)
<p2 category>=adjective
<p2 qualitative>=true
<p2 relational>=false.
```

Figure 6 Example of Meta-Rule used for disambiguation of adjectives

This rule indicates that any adjectives directly preceded by a grading adverb has a qualitative function. If the source part of the Meta-Rule subsumes and can be unified with feature structures representing a sequence of words, these structures will be replaced by the target part of the Meta-Rule. In this example, features *<qualitative>=true* and *<relational>=false* are added to the adjective.

This Meta-Rule will correctly assign a qualitative function to the adjective *tendu* in the sequence "*un climat de plus en plus tendu*" *(a situation more and more tense)*, witch can be minimally represented by the following sequence of feature structures :

<form> = un
<lemma> = un
<category> = determiner
<type> = particle
<defined> = false.

<form> = climat
<lemma> = climat
<category> = noun
<type> = common.

<form> = de_plus_en_plus
<lemma> = de_plus_en_plus
<category> = adverb
<type> = grading.

<form> = tendu
<lemma> = tendu
<category> = adjective
<qualitative>= true
<relational>= true.

In this case, the value of the feature *<relational>* will be changed to *false* by the Meta-Rule. However this rule is not aimed at recognizing the qualitative function of the adjective *dangereux* in "*un climat dangereux à tous égards" (a dangerous situation in all respect).*

<form> = un
<lemma> = un
<category> = determiner
<type> = particle

<*defined*> = *false.*

<*form*> = *climat*
<*lemma*> = *climat*
<*category*> = *noun*
<*type*> = *common.*

<*form*> = *dangereux*
<*lemma*> = *dangereux*
<*category*> = *adjective*
<*qualitative*>= *true*
<*relational*>= *true.*

<*form*>= *à_tous_égards*
<*lemma*>= *à_tous_égards*
<*category*>= *adverb*
<*type*>= *general* .

This sequence of feature structures will remain unchanged by the Meta-Rule (and the function of the adjective ambiguous) because of the mismatch between the postposition of the adverb and the value of its feature <*type*> which value is not equal to *grading*.

6.3 Building of a reference corpus

For this comparison between a fixed list approach and the rule-based approach, we used a sample of 13 papers from *Le Monde* taken from the PAROLE corpus. These texts has been extracted from the *Economy* section of the newspaper and represent around 10.000 words. The corpus was first tagged using the CORDIAL[8] tagger and then converted into feature structures. A refinement of the original tag set has been provided by adding an information via Meta-Rules about grading adverbs for 129 frequent adverbs. Thus, adverbs such has *"très" (very)*, *"plus" (more)* or *"extrêmement" (extremely)* gained a new feature <*type*>= *grading* which was not present after the original tagging made by CORDIAL and its conversion into feature structures. The next step was a manual categorization of adjectives between relational and qualitative. No adjectives have been left ambiguous. A rectification of the data was necessary in order to correct the errors made by the POS tagger. This is commonly the case of verbs erroneously analysed as adjectives or a bad tokenisation ("*Ministre de l'Intérieur*" *(Minister of the Interior)* recognized as a single token but "*Premier Ministre*" *(Prime Minister)* identified as an Adjective followed by a Noun). After this correction, the corpus contained 507 adjective occurrences with 378 qualitative for 129 relational uses.

6.4 Results and discussion

This corrected corpus serves as reference to evaluate and compare the two approaches for adjective categorization. At the beginning of the experience a version of the corpus was created, where all adjectives were ambiguous (the values of their features <*qualitative*> and <*relational*> were both *true*). The goal of the evaluation is to measure the recall and precision provided by the different methods for adjective marking. By recall, we mean the ratio of adjectives which have been disambiguated *((#total - #ambiguous)/ #total * 100)*, correctly or not, while precision indicates the ratio of well-tagged adjectives after disambiguation *(#well-tagged / (#total - #ambiguous) * 100)*. Three tests have been carried out using respectively the plain adjective list described above, the set of Meta-Rules and a mix of list and rules on the ambiguous version of the corpus. Results are compared against the reference corpus in order to determine which of these approaches is the most efficient to distinguish between qualitative and relational uses of the adjectives. Figure 7 shows the results obtained.

[8]See http://www.synapse-fr.com.

	List	Rules	List + Rules
Correct	222	178	290
Wrong	6	0	6
Ambiguous	278	328	210
Total	506		

Recall (%)	45.05	35.17	58.49
Precision (%)	97.36	100	97.97

Figure 7 Compared results for relational / qualitative tagging

The values in the row *Correct* indicate how many adjectives has been correctly categorized as whether qualitative or relational according to the different methods. *Wrong* gives the number of wrong marked adjectives, while the numbers in *Ambiguous* refer to the occurrences of adjectives not covered neither by the list nor by the Meta-Rules.

Using only the contextual Meta-Rules improves the precision of the adjective categorization compared to the use of the fixed list but provides a loss of recall that reaches 10%. The best solution seems to be a mixed use of the list and the rules, which improves the recall with a relatively equal rate of precision. However this gain in recall is relatively moderate. It could be explained by a partial overlap of coverage between the two approaches (the adjectives recognized are often the same).

This example of linguistic feature marking (a new feature is added to the description of a corpus) illustrates the use of feature structures as a representation format. The combined use of fixed information and contextual rules improved the realization of such a subtle opposition as the one between relational and qualitative for adjectives, compared to the projection of a lexicon alone. This method enables the characterization of individual word occurrences, rather than word types, without requiring an important learning phase (Hatzivassiloglou, 2000). Another aspect is the fact that this mixed method allows non-usual cases to be marked correctly, like adjectives which have in context a different function than their most probable one (ex: "*a very Parisian atmosphere*"). However the results of this experience could be surely improved by refining the content of the fixed list and of the rules. Some semantic information would be interesting to solve the ambiguity between relational and qualitative functions of adjectives (this is required to disambiguate the example with "*économique*" provided at the beginning of this section), as well as regular expression operators in Meta-Rules in order to use morphological information (word endings).

7 Conclusions

In this paper we report on the use of feature structures to represent the linguistic information of a corpus. Experience of corpora profiling in the TyPTex project shows that this approach helps to represent any kind of linguistic information, independently of the tools used for tagging. Feature structures is an unifying format which can be used to map from an annotation scheme to an other, in the spirit of Atwell et al. (1994).

Feature structures formalism also helps to handle a set of features with Meta-Rules based on unification. We showed that corpora exploration requires to modify the linguistic features in order to obtain new results and thus to change the point of view on the data. Another aspect is that the features available at the output of a POS tagger may not be sufficient for a given experimentation, one needs to add some information to get the awaited level of marking. By defining Meta-Rules to operate on the feature structures representing a corpus, one can modify the information in a contextual way. An example of a mixed approach between projection of information from a static list and contextual marking via Meta-Rules showed that feature structures can improve the reliability and coverage of the marking.

Acknowledgements

The authors wish to thank Marianne Dabbadie and Lee Humphrey (LexiQuest) for their help during the writing of this paper.

References

Adda G, Mariani J, Lecomte J, Paroubek P, Rajman M 1998 The GRACE French Part-Of-Speech Tagging Evaluation Task. In *Proceedings of LREC'98 (1st International Conference on Language Resources and Evaluation)*, Granada, Spain, pp 2433-2441.

Atwell E, Hughes J, Souter C 1994 AMALGAM: Automatic Mapping Among Lexico-Grammatical Annotation Models. In Klavans J, Resnik P (eds), *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*. Las Cruces, Association for Computational Linguistics, pp 11-21.

Bartning I, Noailly M 1993 Du relationnel au qualificatif: flux et reflux. In *L'information grammaticale* 58(1): 27-32.

Biber D 1988 *Variation across speech and writing*. Cambridge, Cambridge University Press.

Biber D 1993 Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2): 243-258.

Biber D 1995 *Dimensions of register variation : a cross-linguistic comparison*. Cambridge, Cambridge University Press.

Bird S, Liberman M 2001 A formal framework for linguistic annotation. *Speech Communication* 33(1): 23-60.

Fellbaum C, Gross D, Miller K 1990 Adjectives in WordNet. *International Journal of Lexicography* 3(4): 265-277.

Folch H, Heiden S, Habert B, Fleury S, Illouz G, Lafon P, Nioche J, Prévost S 2000 TyPTex: Inductive typological text classification analysis for NLP systems tuning/evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, pp 141-148.

Gazdar G, Pullum G, Carpenter R, Klein E, Hukari T E, Levine R D 1988 Category structures. *Computational Linguistics*, 14(1): 1-19.

Habert B, Salem A 1995 L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *TAL*, 36(1-2): 249-276.

Hatzivassiloglou V, Wiebe J 2000 Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, pp 299-305.

Illouz G, Habert B, Fleury S, Folch H, Heiden S, Lafon P 1999 Maîtriser les déluges de données hétérogènes. In Condamines A, Fabre C, Péry-Woodley M-P (eds), *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Cargèse, pp 37-46.

Illouz G, Habert B, Folch H, Fleury S, Heiden S, Lafon P, Prévost S 2000 TyPTex: Generic features for Text Profiler. In *Content-Based Multimedia Information Access (RIAO'00)*, Paris, pp 1526-1540.

Illouz G 2000, Typage de données textuelles et adaptation des traitements linguistiques. Doctorat d'informatique, Université Paris-Sud.

Ingenia 1995 *Manuel de développement Sylex-Base*. Paris, Ingenia – Langage naturel.

Jacquemin C 1994a FASTR: A unification-based front-end to automatic indexing. In *Proceedings of Intelligent Multimedia Information Retrieval Systems and Management (RIAO'94)*, New York, pp 34-47.

Jacquemin C 1994b FASTR: A unification grammar and a parser for terminology extraction from large corpora. In *Proceedings of Journées IA'94*, Paris, pp 155-164.

Jacquemin C 1997 *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes.

McEnery A, Wilson A 1996 *Corpus linguistics*. Edinburgh, Edinburgh University Press.

Miller K 1998 Modifiers in WordNet. In Fellbaum C (ed), *WordNet: an electronic lexical database*. Cambridge, MIT Press, pp 47-67.

Ruch P, Gaudinat A 2000 Comparing corpora and lexical ambiguity. In *Proceedings of the Comparing Corpora Workshop (38th Annual Meeting of the Association for Computational Linguistics)*, HongKong, pp 14-20.

Sekine S 1998 The domain dependence of parsing. In *Proceeding of the Fifth Conference on Applied Natural Language Processing (Association for Computational Linguistics)*, Washington, pp 96-102.

Shieber S 1985 *An introduction to unification-Based Approaches to Grammar*. Stanford, CSLI Lecture Notes 4 (Center for the Study of Language and Information).

Tambouratzis G, Markantonatou S, Hairetakis N, Vassilliou M, Tambouratzis D, Carayannis G 2000 Discriminating the registers and styles in the Modern Greek language. In *Proceedings of the Comparing Corpora Workshop (38th Annual Meeting of the Association for Computational Linguistics)*, HongKong, pp 35-43.