Seminar:
Korpusbasierte englische Sprachwissenschaft
(English Corpus Linguistics)
Dr. Elke Teich
Winter Semester 2001/2002

Interaction between Tag Set Design and Multilingual Information Extraction

Student: Garance PARIS
19th March, 2002

English Linguistics and Translation Studies
Department 4.6 Applied Linguistics, Translation and Interpreting
University of the Saarland

Summary

In this paper, we first take a look at why the part-of-speech annotation in the subcorpora of a multilingual corpus is often incomparable. We detail a number of sources of non-equivalence in the PoS-annotation, which are variously due to the tag set, the tagging scheme or the tagger itself: syntactic properties of languages, differing granularity, interaction between linguistic criteria and computational tractability, divergence and vagueness of linguistic theories, inclusion of semantic information, tokenisation difficulties, tagging "in context" or not, and tagger accuracy.

We then examine how this could be remedied by developing or adapting the tag sets with multilingual information extraction in mind, and suggest some tag set characteristics which tend to make the PoS-annotation more comparable: medium granularity, hierarchicality, and international guideline compliance (cf. EAGLES and MULTEXT, Leech and Wilson, 1996, Ide and Véronis, 1994, Calzolari and Monachini, 1996).

Finally, we also consider what can be done to retrieve syntactic information from the corpus in spite of non-equivalence in the PoS-annotation, and explore three different ways of retrieving syntactic structures from the corpus when the PoS-annotation cannot be relied on: extraction based on specific word-forms, extraction using parallel concordances, and extraction requiring parsing.

Contents

1. Introduction

Corpora are an invaluable source of knowledge in linguistics: whereas most research on grammar and syntax was previously based on the researcher's intuition about language, today the existence of large corpora and computer-based methods to analyse them make it possible to study syntactic structure as it is actually produced by people in everyday life.

Usually in corpus linguistics, studying a particular syntactic phenomenon involves finding its concrete instances in a corpus and inspecting the results. Sometimes however, the structures we are interested in cannot be extracted straightforwardly from the corpus, and in such cases, it is often helpful to add some kind of information to the corpus, information which in turn will allow us to get more out of the corpus.

A French example of such a case would be trying to extract from a corpus all sentences containing an auxiliary use of *être* (Engl. "to be"). Knowing that in French, all compound tenses involve the past participle, we might attempt to formulate this query as searching for sequences composed of a form of *être* and a past participle.

However, we would then hit upon the difficulty that there is no way to tell from the raw data which words in the corpus are past participles: although by far the majority of participles (those of the first conjugation) end in *-é(e)(s)*[1], there are also many participles (from other conjugation classes or irregular) which would not be found if we defined a participle in our query on the basis of these sole endings, and there are also many words which end in *-é(e)(s)* but are not participles.

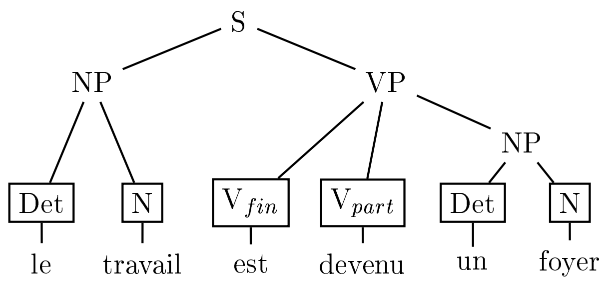
Using a terminology which is common in information retrieval, we would say that the first case illustrates suboptimal **recall** (not all auxiliary-past participle sequences are found), whereas in the second case, it the degree of **precision** which is suboptimal, since not all sequences in which *être* is followed by a word ending in *-é(e)(s)* denote an auxiliary use of *être* (Witten and Frank, 2000).

In this example, as in many other cases, the necessary information can be provided by labelling each word in the corpus with part-of-speech (PoS) information. Additionally, depending on the part-of-speech category, it is often useful to add some basic information on the inflected form: in the above example, this will enable us to distinguish between finite forms of verbs, participles and infinitives.

Formally, this amounts to specifying the preterminal nodes of a parse tree (Leech, 1997), thus allowing us to use information on both terminal and preterminal nodes in our queries, instead of being limited to queries based on the leaves only. Figure 1 illustrates this, and shows how it can help us solve the above problem: in the sentence *Le travail est*

[1] i. e. in *-é*, *-és*, *-ée* or *-ées*, which are the respective endings of the masculine singular, masculine plural, feminine singular and feminine plural forms of the past participle of verbs from the first conjugation class.

Figure 1: PoS-tagging allows the retrieval of French participles



devenu un foyer (Engl. "The workplace has become a home."), the word *devenu* (Engl. "become") is now recognisable as a past participle, although *devenir* is an irregular verb whose past participle does not end in *-é*.

However, if the subject of our research requires querying a multilingual corpus, perhaps because we ask ourselves how specific syntactic constructions are rendered in translations from one language into another, or for any other research goal, then in some cases formulating queries based on the PoS-annotation may turn out to be non-trivial. Inevitably, there will be differences in the way each subcorpus is marked up with PoS-information, differences which may be due

- to varying choices of tags to represent the PoS-information, i. e. to differences in the **tag sets**;
- to different ways of assigning tags to words, i. e. to differences in the so-called **tagging scheme**;
- or to varying accuracies of the tagging software, the **tagger**, used for the automatic annotation.

Plan of the Paper

In order to understand why and how each of these factors influences the PoS-annotation, we will first give a brief introduction to the functioning of programmes for automatic PoS-tagging (chapter 2). Following this, we will then examine in detail the sources of differences between tag sets, tagging schemes and taggers (chapter 3). This will then bring us to ask ourselves how this can be remedied: how should tag sets be devised in order to ensure that we can later extract equivalent information from the different subcorpora of a multilingual corpus (chapter 4)? Finally, we will also take a look at what can be done when the PoS-annotation of the subcorpora of a multilingual corpus is not comparable, and explore a few approaches which may allow us to nonetheless extract the information we are interested in (chapter 5).

2. Automatic PoS-Tagging

Part-of-speech annotation is often done by software tools called part-of-speech taggers, whose job it is to automatically assign to each word of a text a tag indicating its part-of-speech, possibly complemented by some additional morphological information.

This is not a trivial task, for opening any dictionary reveals that most words can belong to more than one part-of-speech category. Thus the tagger's job is to determine which of these categories is the most likely one in a given sentence (Manning and Schütze, 1999, chap. 10), a task which some taggers achieve better than others.

The accuracy of the result, however, does not depend only on the tagging software: using knowledge about how tagging works also allows to design tag sets with the objective of obtaining better results, but unfortunately, this is often done at the expense of comparability of tag sets across languages.

In order to enable us to spot such unfortunate changes to a tag set, this chapter will explain the basic ideas which underlie automatic taggers.

2.1. Sources of Information for Automatic Taggers

There are various approaches to automatic part-of-speech tagging, which diverge in the mathematical methods they implement. But whatever the formal approach, one thing is common to all taggers: they are all based on the same linguistic information, which is of two kinds (Manning and Schütze, 1999, chap. 10): **lexical information**, which predicts possible tags based on the word concerned, and **distributional information** (also sometimes termed 'structural' or 'syntagmatic information'), which builds on the fact that some sequences of tags are common and others extremely unlikely or impossible, thus deriving constraints on the PoS-category of a word from its immediate neighbourhood.

This linguistic information is usually contained in separate language and tag set-dependent files, which are distinct from the tagging engine itself, and is gathered during a so-called **training** phase, during which a learning algorithm observes in a previously (hand-)tagged corpus the regularities in the tag sequences, and uses them to deduce the information which the tagger needs.

In order for the tagger to be reliable, the size of this training corpus should be adapted to the granularity level of the linguistic model to be learnt: learning a very fine-grained model, i. e. one which makes very fine distinctions between parts-of-speech or one which attempts to model syntactic dependencies over more than just a few words, requires a much larger training corpus than learning a coarser model.

Consequently, since the corpora available for training are often of limited size, the extent to which taggers take distributional information into account is actually quite limited. Present-day taggers ignore anything but short-distance syntactic dependencies

(Garside and Smith, 1997): probabilistic or rule-based approaches alike use, in the best of cases, information on the preceding or the two preceding tags, but never more, because the construction of more fine-grained models would require too large corpora (Manning and Schütze, chap. 10). As a result, taggers are unable to make some kinds of linguistic distinctions—we will presently see an example in section 2.2.

In the context of multilingual tagging, it is also interesting to note that the degree to which lexical and distributional information each contribute to the tagging process is language-dependent. Indeed, language typology shows the degree of word order freedom in a language to be proportional to the amount of inflectional morphology. For this reason, in languages with a rather free word order, but a rich inflectional morphology, it is lexical information which will be most reliable, whereas in languages with a very rigid word order and little inflection, the informativeness of distributional information will be greater, so that all in all, similar accuracies are achieved by taggers for different languages (Manning and Schütze, chap. 10).

2.2. An Example Using a Probabilistic Tagger

Let us now use the TreeTagger (Schmid, 1994, 1995, Stein and Schmid, 1995) to illustrate the interplay of both sources of information that taggers draw on. The engine of the TreeTagger is language independent; information relative to each language is stored in parameter files,[2] under the form of probabilities: the lexical information is represented by associating words with their possible tags in order of frequency, and the distributional information is represented by a decision tree, in which the probability that a given word should be attributed some tag depends on a variable number of tags preceding it.

As an example, take the French word *neige* (Engl. “snow”). When we ask for the lexical information associated with it,[3] the TreeTagger tells us that this word can be a noun (`NOM`) and a verb in the indicative present (`VER:pres`) or in the subjunctive present (`VER:subp`), but that the noun reading is predominant:

```
neige NOM 0.990 VER:pres 0.009 VER:subp 0.000
```

However, the linguistic knowledge the tagger has about the way PoS-categories are distributed across French sentences also tell it that after a personal pronoun (`PRO:pers:conj`) beginning a sentence, the next word is more likely to be a verb in the indicative present than a verb in the subjunctive or a noun:[4]

```
0.110331        VER:pres
0.020060        VER:subp
0.001003        NOM
```

[2] The examples here use the Linux version of the TreeTagger. In some cases, the Unix version of the parameters return different values.

[3] This information was obtained with the following command, where `-proto-with-prob` means “For each word, print lexical information with tag probabilities”:

```
$ echo 'neige' | tree-tagger $TTPARAM/french.par -token -proto-with-prob
```

Using both the lexical and the distributional information, the algorithm used by the TreeTagger thus computes the following probabilities for *Il neige!* (Engl. "It's snowing!"), which allow us to conclude that in this context, *neige* is most likely to be a verb in the present indicative:[5]

```
il            PRO:pers:conj 1.000000
neige         VER:pres 0.955877 VER:subp 0.038536 NOM 0.005587
```

Now, remember how we said in section 2.1 that automatic taggers are unable to make out long-distance syntactic dependencies. An example of a distinction which an automatic tagging tool cannot make is between French indicative and subjunctive verbs of the first conjugation, because the verb forms are ambiguous. Thus in *"Comme je voudrais qu'il neige!"* (Engl. "I wish it would snow!"), although the main verb *voudrais* requires the verb of the following subordinate clause to be in the subjunctive, the tagger, which cannot understand a dependency going back three words, does not assign the correct tag, which should be `VER:subp` (verb in the subjunctive present):

```
qu'     CON:sub
il      PRO:pers:conj
neige   VER:pres
```

[4] This data was gathered by observing the transition probability tree, which can be output by the command `tree-tagger $TTPARAM/french.par -print-prob-tree`. The relevant part of the tree, i. e. the section which gives the probabilities of tags following a personal pronoun at the start of a sentence, was saved in the file `PRO+1`. The command line then was:

```
$ less PRO+1 | awk '{print $2" "$1}' | sort -nr | egrep "VER:pres|NOM|VER:subp"
```

[5] The command line here was:

```
$ echo 'Il neige!' | tokenize-french | tree-tagger $TTPARAM/french.par -token
-threshold 0.000001 -prob
```

3. Sources of Non-Equivalence in the PoS-Annotation

Unfortunately, when we set out to tag corpora in different languages, frequently we have to use resources (tagging software, tag sets,...) which have been developed independently by different people, who did not consider multilingual applications, so that there arise some questions as to the extent of the comparability of the annotation resulting from the tagging process.

In order to illustrate this, we will take the case of a French-German parallel corpus tagged with the TreeTagger (Schmid, 1994, 1995, Stein and Schmid, 1995) using the French and German parameter files, which respectively use a slightly modified version of the tag set developed by Stein (1995) and the Stuttgart-Tübingen tag set (abbreviated STTS; Schiller et al., 1995).

Let us now take a look at the possible impact of the tag sets, tagging schemes and accuracy of the tagger on the comparability of the PoS-information in the French and German subcorpora of such a corpus.

3.1. The Tag Sets

Although the core of a tag set will always be the ten or so traditional major word categories and their subcategories, even within any one language, available tag sets differ greatly in both the number of their tags and in the tags themselves: for English, Leech (1997) gives as indication that tag sets can have anywhere between 30 to 200 members, and adds that tag set size tends to be higher in languages that use more inflection.

The two tag sets we will compare here, the Stein (1995) tag set and the Stuttgart-Tübingen tag set (Schiller et al., 1995), are built on the same ten major part-of-speech categories, with the addition of a category for particles in German. Their size is roughly similar: the Stuttgart-Tübingen tag set contains 54 tags, of which 48 are PoS-tags and six are devoted to other categories (punctuation, words from foreign languages, truncated words, formulae and other non-words), Stein's tag set has 62 comprising five extra (punctuation, non-words and abbreviations).

Let us first take a look at the tag sets themselves, which are presented in the synoptic table on pages 16–18, and compile a list of the essential differences that can be identified between the two:

1. The STTS does not make a distinction between the definite and the indefinite articles;
2. Stein's tag set has an additional tag for the French partitive article;
3. The correspondence between <PIAT>, <PIDAT> and French equivalents is unclear;

4. Stein's tag set uses <PRO:poss> for attributive possessive pronouns only: there is no "pred"/"attr" distinction as with other pronouns, and thus no equivalent for the STTS's <PPOSS> (instead *[le] mien* is tagged as an adjective);
5. The STTS distinguishes between different kinds of relative pronouns which have different distributions: attributive (*dessen*: <PRELAT>), adverbial (*wo*, *warum*), and those replacing a subject or an object (*der*: <PRELS>). It similarly distinguishes between three types of interrogative pronouns: <PWS>, <PWAT>, <PWAV>. Building on this distinction and realising that adverbial relative pronouns and interrogative pronouns have a similar distribution, Schiller et al. then collapse these classes into one: <PWAV>;
6. The French *y* and *en*, tagged <PRO:clit> in Stein's tag set, only partially correspond to the STTS's <PAV>;
7. Whereas the STTS does not mark case and person of personal pronouns, Stein's tag set has information on person and conjoint, disjoint, and clitic forms of French pronouns;
8. Similarly, the STTS does not mark the person of reflexive pronouns, but Stein's tag set does;
9. The STTS makes a distinction between full verbs, modals and auxiliaries, Stein's tag set does not (although it is theoretically unclear whether a category for modal verbs is relevant for French);
10. VVIZU is not a relevant category for French;
11. Stein's tag set has a tag for the French present participle, but the STTS tags German present participles "in context", i. e. depending on the way they function in a sentence;
12. Stein's tag set has no tag for imperative;
13. Stein tags tense, mode and person of finite verbs, the STTS does not;
14. The STTS distinguishes between predicative and attributive adjectives, but not Stein's tag set;
15. The STTS has a number of additional categories which are relevant only for German: postpositions, circumpositions, particles, <TRUNC>;
16. Stein's tag set has no tag for foreign words and none for "answer particles";
17. Stein's tag set does not have a tag for conjunctions introducing an infinitive subordinate clause, although this would be relevant for French;
18. The STTS has a tag for *zu* preceding infinitives, but Stein's tag set always tags *de* as a preposition, also in this case;
19. The STTS has an extra tag for comparative conjunctions in German, but the French comparative *que* is collapsed with other subordinating conjunctions in Stein's tag set;
20. Abbreviations are tagged "in context" in the STTS, but have a separate tag in Stein's tag set.

Now that we know what differences exist between the two tag sets we are considering, our next question should be: why do these differences exist? Tag set differences have a number of reasons, some of which some of which may occur between any two tag sets, even within the same language, and some of which arise only in a multilingual perspective. We will now explore some of these sources of non-equivalence, each of which is illustrated by some of the differences in the list above.

3.1.1 Syntactic Differences (2, 3, 6, 7, 9, 10, 11, 15). Of course, each language has a syntax of its own. Thus try as we might, there will be differences between tag sets for different languages, reflecting their syntactic peculiarities.

In some cases it may be so that some categories which are relevant for one language find no correspondence in another: an example of this would be the tag for the French partitive article in Stein's tag set, a category irrelevant for German, or the distinction of modal verbs in the STTS, since modals behave differently from full verbs in German, but not in French. In other cases, it can happen that a correspondence is only partial: although the French *y* and *en* often correspond to pronominal adverbs in German, not all pronominal adverbs in German correspond to *y* or *en* in French.

3.1.2 Granularity (1, 4, 5, 7, 8, 9, 12, 13, 14, 16, 17, 19). Some tag sets make more subtle distinctions within one category than others, marking for instance the subcategorisation frames of verbs, more morphological inflection, or distinctions depending on the distribution in sentences (see also 3.1.3 below).

Such differences can be observed in several places in the STTS and Stein's tag set, although both tag sets are of roughly similar size. For example, the STTS is more precise when it comes to relative and indefinite pronouns, but Stein's tag set is more precise in that it marks more morphological information because French has more verbal inflection, and in that it marks person and form of French personal pronouns, whereas the STTS does not mark the case of German personal pronouns.

In addition, in some cases in which a tag set has been developed only on limited training material, it may also happen that some categories are simply absent from the tag set, such as imperative in Stein's tag set.

3.1.3 Desire for Accuracy with Automatic Taggers (5, 9, 14). Not only "sound" linguistic criteria have a say in tag set design, but also practical criteria: there is often a trade-off between, on the one hand, the linguistic quality of a tag set, also termed **external validity** because it refers to "the extent to which [the tag set] allows retrieval of all important grammatical distinctions of the language" (Sampson, 1995, cited according to Déjean, 2000) in subsequent linguistic processing, and, on the other hand, its computational tractability, or **internal validity**, meaning "the extent to which [its tags are] useful in aiding the disambiguation process, and increasing the accuracy of tagging" (ibid.; see also Leech, 1997).

In particular, Teufel et al. (1996) prove that the distributional information mentioned in section 2.1 can be used to increase the internal validity of a tag set: since part of the information that taggers rely on incorporates knowledge on the way part-of-speech categories are distributed across sentences, the developers of the STTS came up with the idea of attempting to obtain a higher accuracy by both collapsing categories which have the same distribution and introducing different tags for the same part-of-speech category appearing in different contexts.

This explains why the STTS differentiates between attributive and predicative adjectives (one PoS-category occurring in different contexts has been split into two tags), and also how it came to have one tag, <PWAV>, for what corresponds to a part of the relative pronouns and a part of the indefinite pronouns in Stein's tag set.

Indeed, due to the fact that the linguistic information the tagger has at its disposal takes only the one or two preceding tags into account, both adverbial relative pronouns (as in *der Ort* <N> *(...), wo* <PWAV> *er wohnt*, Engl. "the place where he lives") and adverbial interrogative pronouns (as in *sie fragt* <V> *(...), wo* <PWAV> *er wohnt*, Engl. "she asks where he lives") appear to the tagger as having the same distribution. Since differentiating reliably between the two would require information on the attachment site (verb or noun) of the pronoun, but the verb/noun and pronoun are often separated by several words in real-life text, this is a difficult distinction for a tagger to make, and consequently the developers of the STTS decided to merge these two categories with similar distributions.

However, although the splitting of a category is most often unproblematic because the subcategories can easily be merged again later, it cannot be said that collapsing categories with similar distributions is without consequences for multilingual PoS-tagging, because the distribution of categories is unlikely to be the same in different languages.

3.1.4 Linguistic Theories (3, 18). Sometimes the underlying linguistic theories, too, can be responsible for differences in tag sets, in such areas where there is no clear cut theory as to what words should belong to which category. Véronis and Khouri (1995) mention this problem, and write that it is especially acute for closed-class words such a determiners, pronouns, the French 'indefinite adjectives'[6], etc. Again, some differences between the STTS and Stein's tag set have their origin here, for instance the unclear correspondence between <PIAT>, <PIDAT> and <PRO:indef:attr> for attributive indefinite pronouns.

But even when the theoretical positions are well defined, it can happen that not all theories agree on the extension of some PoS-categories: for multilingual tagging in particular, different traditions in the study of linguistics must also be taken into consideration, such as those of Romance languages and Germanic languages.

An example here would be the distinction between 'attributive' and 'predicative' pronouns, on which the STTS and Stein's tag set agree and which was taken up in the EAGLES recommendations (Leech and Wilson, 1996, see chapter 4), but which is unusual in French linguistics. Indeed, other tag sets commonly used in France (Rajman et al., 1997, Abeillé et al., 2001) classify only 'predicative pronouns' as pronouns and considers the 'attributive pronouns' to be determiners, a stance which was also followed in the MULTEXT project (Ide and Véronis, 1994, see chapter 4), which thereby broke with the EAGLES recommendations.

Moreover, in addition to these sources of differences identified by comparing the French and German tag sets used by the TreeTagger, at least one more is mentioned by Véronis and Khouri (1995):

[6] French linguistics commonly group under the category 'indefinite adjective' a few words among which *autre*, *quelconque* and *même*, which are most often adjectives but may also function as determiners in some set expressions or in archaïc and literary text (Grévisse and Goosse, 1989, paragraph 246).

3.1.5 Inclusion of Semantic Information. In addition to PoS-information, some tag sets also include some lexical semantics in their categories, be it for the needs of a specific application such as information retrieval (Véronis and Khouri, 1995, give as example the tag set of IBM France which distinguishes between proper nouns for countries, cities, persons, etc.), or in order to increase accuracy.

An illustration of the last case is given by Déjean (2000), who explains that in English, for example, temporal nouns such as *today*, *tomorrow*, *tonight* and *yesterday* may also function adverbially at the clause level, and thus distribute differently from other nouns, so that introducing a separate category for them can prove helpful (see chapter 2), as was done in the Susanne tag set (Garside et al., 1987, cited by Déjean).

3.2. The Tagging Scheme

By itself however, the tag set is not enough to describe a tagging project: there are many more decisions involved, the description of which builds the **tagging scheme**. Two aspects of tagging schemes will be considered here because they have an influence on the multilingual comparability of the resulting PoS-annotation: the way in which a corpus is divided into units to be tagged, a step called **tokenisation**, and the way in which tags are assigned to words.

3.2.1. Tokenisation

At first glance, tokenisation may seem a trivial task: in the majority of cases, we can assume that the units are orthographic words delimited by whitespace or punctuation, but there are some deviations from this, the main two being multiwords and mergers (Leech, 1997), which, from the perspective of multilingual PoS-annotation, both tend to disrupt a desirable one-to-one correspondence between words in different languages.

Multiwords

Multiwords are sequences made up of more than one orthographic words which for tagging purposes would ideally need to be treated as one unit. From the point of view of multilingual PoS-tagging, it is especially important that multiwords belonging to closed-class words, which often indicate grammatical relationships, receive appropriate treatment, because what is a multiword in one language frequently corresponds to only one word in another language.

A case of ideal treatment of multiwords is the TALaNa annotated corpus, which was entirely tagged by hand. It's tagging scheme (Abeillé and Clément, 1999) supplied detailed graphical, morphological, syntactic and semantic criteria for the identification of multiwords, and then required that the annotators join the parts of a multiword with underscores (as in *au_cours_de*, German "während", Engl. "during"; Abeillé et al., 2001), the resulting unit then being tagged with a single part-of-speech.

However, although this is a comparatively easy treatment for human annotators to apply, it is not possible to do the same with presently available methods when the PoS-annotation of a corpus is to be done automatically.

For this reason, in the Stuttgart-Tübingen tagging scheme, which was designed for use with automatic PoS-taggers, multiwords are instead tagged as if each part were independent from the others: *vor* <APPR> *kurzem* <ADJA> (preposition + attributive adjective; French "récemment", Engl. "recently"), *so* <ADV> *daß* <KOUS> (adverb + subordinating conjunction; French "de sorte que", Engl. "so that").

A slightly better approach applicable to automatic tagging, however, is the one chosen by Stein in the French parameters for the TreeTagger.[7] Indeed, although Stein also tags most of grammatical multiwords as if they were composed of independent words, as in *tel* <ADJ> *que* <CON:sub> (Engl. "such as", German "wie"), he also introduces numbering of the parts for the most frequent French multiwords, thus allowing them to be identified as one unit: *alors* <CON:sub:1st> *que* <CON:sub:2nd> (subordinating conjunction; Engl. "although").

Mergers

The second problem for tokenisation are mergers, in which one orthographic word corresponds to more than one morphosyntactic word. From the point of view of automatic processing, three categories can be distinguished:[8]

1. The mergers of the first category are those in which two words are written in one, possibly with one of them being phonologically reduced, but there is some clear separation marker at the junction, such as an apostrophe or a hyphen, as in the German *geht's* (=*geht es*) or the French *c'est* (=*cela est*).

   Although this category is actually the easiest to deal with, because the two parts of the merger simply need to be separated at the junction marker, the separation may still require fairly involved tokenisation rules, because not all hyphens and apostrophes indicate mergers.

   For example, in French, the general rule applied by the TreeTagger when it encounters a hyphen is to treat the entire word as one unit, but there are exceptions, such as *-mêmes*, which is usually separated from the following word, except when it is preceded by a personal pronoun, thus allowing to identify reciprocal pronouns: e. g. *lui-même* is thus an exception to the exception.

2. The second category is a bit more difficult to deal with, because there is no trace left of the junction, both words being simply concatenated. In some cases, phonological reduction may also take place, and sometimes be marked by an apostrophe

[7] Unfortunately, no document describing the tagging scheme is available for Stein's French parameter files for TreeTagger, so that for the most part, the information presented here comes from observing the output of the French TreeTagger and from taking a look at the supplied scripts.

[8] Leech (1997) proposes another classification, which however is less useful for our purposes. His categories are the following: mergers involving clitics (words which are phonologically reduced and attached to the beginning or the end of another word, as in *c'est* or *haven't*), purely orthographic mergers (without reduction; e. g. Italian *vendetelo*, verb + object pronoun), mergers in which the blending is accompanied by a modification of the original words, thus making separation of the components very difficult (corresponding to our third category), and compound words, which we do not mention because automatic taggers usually deal with them in a trivial one unit-one tag fashion.

occurring **within** one of the parts: German *zum* (=*zu [de]m*), Engl. *haven't* (=*have n[o]t*).

3. Finally, the last category comprises the mergers which are most difficult to deal with: those in which the blending of both words results in modifications which make it very difficult to separate both parts, as in the French *du* (=*de le*) and *aux* (=*à les*), or in the English *won't* (=*will not*).

In the second category, it is in principle possible to separate both words, which is what the TreeTagger does for the negative forms of English verbs ending in *-n't*: *haven't* for instance is broken up into *has* <VH> (base form of a verb) and *n't* <RB> (adverb; Penn-Treebank tag set, Santorini, 1991). However, this seems to be the single case in which the two parts of a merger built by concatenation are separated, for most tagging schemes choose not to split this type of merger.

Instead, for both the second and third categories, some tagging schemes resort to introducing what is called **compound tags** for both mergers built by concatenation and mergers built by blending. These compound tags are actually extensions of the basic tag set which are created by concatenating two basic tags, and thus offer the advantage of allowing the retrieval of both "subtags" by using wildcards in a query. The Stuttgart-Tübingen tagging scheme, for example, introduces <APPRART> for German concatenated mergers such as *zum*, thus allowing to retrieve all prepositions including *zum* with `APPR*` or all articles with `*ART`, by using the 'Kleene star' to replace any combination of characters.

As for the French parameters for the TreeTagger, they proceed similarly in some cases, but not in all: Stein introduce <PRE:det> for the French preposition-article blends *du*, *aux*, etc., but has no compound tag for preposition-relative pronoun blends such as *auquel* (=*à lequel*) and *duquel* (=*de lequel*), which instead are tagged only as <PRO:rela>, resulting in information loss.

3.2.2. Context-Dependent vs Context-Independent Tagging

In addition to tokenisation, the way in which tags are assigned to words by a tagging scheme is also likely to influence the comparability of the part-of-speech annotation in different languages.

In their definition of PoS-tagging, Garside and Smith (1997) insist that it should be done "in context", meaning that a word is tagged to reflect its actual PoS-category in the text and not the PoS-category it would have in a lexicon. However, in the case of low frequency words, some tag sets prefer to assign context independent tags (Véronis and Khouri, 1995), because these words are unlikely to be encountered during the training of the tagger, so that the tagger may not have the necessary distributional information available, and thus these words could not be tagged reliably with context dependent tags (see examples 11 and 20 in the list in section 3.1).

This is particularly frequent with numerals, and the standards proposed by the EAGLES project (Leech and Wilson, 1996) actually recommend context independent tagging in this case. Thus, although the Stuttgart-Tübingen and Stein's tag sets have corresponding tags for numerals (<CARD> and <ADJ:num:card>), this could introduce a difference in the PoS-annotation should both of these be opposed to a text tagged with a tag set more precise with respect to numerals, such as the one used in the French project GRACE (Rajman et al., 1997) or the German tag set proposed by the MULTEXT project (Calzolari and Monachini, 1996, Ide and Véronis, 1994), both of which tag numerals "in context". This is illustrated by the following example from Rajman et al.:

deux <DET> amis (Engl. "two friends")
les deux <ADJ> amis (Engl. "the two friends")
le chapitre deux <ADJ> (Engl. "the second chapter")
deux <PRO> sont venus (Engl. "two [people] came")
ce soir le programme de la Deux <NOM> est mauvais
(Engl. "this evening the programme on the second channel is is bad")
celui qui a une paire de deux <NOM> ouvre la manche
(Engl. "the player with a pair of two's in his hand starts")

3.3. The Tagger

Finally, the last factor which may influence the comparability of the tagging result is the software tool itself: indeed, depending on the approach of the problem, on the chosen algorithms, on how good a given implementation is, and on whether some kind of post-processing in applied to improve on the basic result, the accuracy of the tagging result can vary considerably.

Nowadays, good taggers achieve between 96 % and 97 % accuracy on newspaper text (which has an average sentence length of over 20 words), which means they make on average one error per sentence (Manning and Schütze, 1999, chap. 10). At first sight, this result may seem a acceptable, but if we take probability laws into account, it is much less encouraging: a tagger which achieves 96 % correctness over words has a rather thin 44.2 % of a chance that it will get all tags in a 20-word sentence right, and a tagger which achieves 97 % fares only slightly better at 54.4 % (compare Manning and Schütze, p. 373).

But most importantly for multilingual information extraction, the problem with automatic taggers is that any tagger-tag set pair has a tendency to systematic errors, thus decreasing the comparability between languages whenever recurring errors on some PoS-category are made for one language but not for the others.

Consequently, for any serious work which involves comparing the results of extracting information from two corpora tagged with different taggers or tag sets, it would be necessary in a preliminary step to determine separately for each language what systematic errors are made, in order to know which PoS-categories cannot be assigned with sufficient reliability.

German (STTS)			French (Stein)	
bestimmter oder unbestimmter Artikel	der, die, das, ein, eine,...	ART	DET:def	Artikel-definit
			DET:indef	Artikel-indefinit
		–	DET:part	Artikel-partitiv
normales Nomen	Tisch, Herr, [das] Reisen	NN	NOM	Nomen
Eigennamen	Hans, Hamburg, HSV	NE	NPR	Eigenname
substituierendes Demonstrativpronomen	dieser, jener	PDS	PRO:demo:pred	Pronomen-demonstrativ (prädikativ)
attribuierendes Demonstrativpronomen	jener [Mensch]	PDAT	PRO:demo:attr	Pronomen-demonstrativ (attributiv)
substituierendes Indefinitpronomen	keiner, viele, man, niemand	PIS ?	PRO:indef:pred	Pronomen-indefinit (prädikativ)
attribuierendes Indefinitpronomen ohne Determiner	kein [Mensch], irgendein [Glas], etliche [Dinge], zuviele [Fragen], etwas [Schokolade]	PIAT ?	PRO:indef:attr	Pronomen-indefinit (attributiv)
attribuierendes Indefinitpronomen mit Determiner	[ein] wenig [Wasser], [die] beiden [Brüder], all [die Bücher], solch [eine Frage], viele [Leute]	PIDAT		
substituierendes Possessivpronomen	meins, deiner	PPOSS	(“[le] mien, [le] sien”:ADJ)	
attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]	PPOSAT	PRO:poss	Pronomen-possessiv
substituierendes Relativpronomen	[der Hund,] der	PRELS	PRO:rela	Pronomen-relativ
attribuierendes Relativpronomen	[der Mann,] dessen [Hund]	PRELAT		
substituierendes Interrogativpronomen	wer, was	PWS	PRO:inter	Pronomen-interrogativ
attribuierendes Interrogativpronomen	welche [Farbe], wessen [Hut]	PWAT		
adverbiales Interrogativ- oder Relativpronomen	warum, wo, wann, worüber, wobei	PWAV		
Pronominaladverb	dafür, dabei, deswegen, trotzdem	PAV	PRO:clit	Pronomen-klitisch-y/en
...			...	
irreflexives Personalpronomen	ich, er, ihm, mich, dir	PPER	PRO:pers:conj:1	Pronomen-personal-unbetont-1.Person
			PRO:pers:conj:2	Pronomen-personal-unbetont-2.Person
			PRO:pers:conj:3	Pronomen-personal-unbetont-3.Person
			PRO:pers:clit:1	Pronomen-personal-klitisch-1.Person
			PRO:pers:clit:2	Pronomen-personal-klitisch-2.Person
			PRO:pers:clit:3	Pronomen-personal-klitisch-3.Person
			PRO:pers:disj:1	Pronomen-personal-betont-1.Person
			PRO:pers:disj:2	Pronomen-personal-betont-2.Person
			PRO:pers:disj:3	Pronomen-personal-betont-3.Person
reflexives Personalpronomen	sich, einander, dich, mir	PRF	PRO:pers:reci:1	Pronomen-personal-reziprok-1.Person
			PRO:pers:reci:2	Pronomen-personal-reziprok-2.Person
			PRO:pers:reci:3	Pronomen-personal-reziprok-3.Person

German (STTS)			French (Stein)	
Infinitiv, voll	gehen, ankommen	VVINF	VER:infi	Verb-Infinitiv
Infinitiv, aux	werden, sein	VAINF		
Infinitiv, modal	wollen	VMINF		
Infinitiv mit “zu”, voll	anzukommen, loszulassen	VVIZU	–	
Partizip Perfekt, voll	gegangen, angekommen	VVPP	VER:pper	Verb-Partizip-Perfekt
Partizip Perfekt, aux	gewesen	VAPP		
Partizip Perfekt, modal	gekonnt, [er hat gehen] können	VMPP		
	([ein schnell] fahrendes [Auto])	(ADJA)	VER:ppre	Verb-Partizip-Praesens
	([die Verfolgung politisch] Andersdenkender)	(NN)		
Imperativ, voll	komm [!]	VVIMP	?	
Imperativ, aux	sei [ruhig !]	VAIMP	?	
finites Verb, voll	[du] gehst, [wir] kommen [an]	VVFIN	VER:cond:1	Verb-conditionnel-1.Person
finites Verb, aux	[du] bist, [wir] werden	VAFIN	VER:cond:2	Verb-conditionnel-2.Person
finites Verb, modal	dürfen	VMFIN	VER:cond:3	Verb-conditionnel-3.Person
			VER:futu:1	Verb-futur-1.Person
			VER:futu:2	Verb-futur-2.Person
		etc.	VER:futu:3	Verb-futur-3.Person
			VER:impf:1	Verb-imparfait-1.Person
			VER:impf:2	Verb-imparfait-2.Person
			VER:impf:3	Verb-imparfait-3.Person
			VER:pres:1	Verb-present-1.Person
			VER:pres:2	Verb-present-2.Person
			VER:pres:3	Verb-present-3.Person
			VER:simp:1	Verb-passesimple-1.Person
			VER:simp:2	Verb-passesimple-2.Person
			VER:simp:3	Verb-passesimple-3.Person
			VER:subi:1	Verb-subjonctif-imparfait-1.Person
			VER:subi:2	Verb-subjonctif-imparfait-2.Person
			VER:subi:3	Verb-subjonctif-imparfait-3.Person
			VER:subp:1	Verb-subjonctif-present-1.Person
			VER:subp:2	Verb-subjonctif-present-2.Person
			VER:subp:3	Verb-subjonctif-present-3.Person

German (STTS)			French (Stein)	
attributives Adjektiv	[das] große [Haus]	ADJA	ADJ	Adjektiv
adverbiales oder prädikatives Adjektiv	[er fährt] schnell, [er ist] schnell	ADJD		
Adverb	schon, bald, doch	ADV	ADV	Adverb
Negationspartikel	nicht	PTKNEG	ADV:neg	Adverb-Negation
Präposition; Zirkumposition links	in [der Stadt], ohne [mich]	APPR	PRE	Praeposition
Präposition mit Artikel	im [Haus], zur [Sache]	APPRART	PRE:det	Praeposition-mit-Artikel-verschmolzen
Postposition	[ihm] zufolge, [der Sache] wegen	APPO	–	
Zirkumposition rechts	[von jetzt] an	APZR	–	
Kardinalzahl	zwei [Männer], [im Jahre] 1994	CARD	ADJ:num:card	Zahl/Zahladjektiv
Fremdsprachliches Material	[Er hat das mit "] A big fish [" übersetzt]	FM	?	
Interjektion	mhm, ach, tja	ITJ	INT	Interjektion
unterordnende Konjunktion mit "zu" und Infinitiv	um [zu leben], anstatt [zu fragen]	KOUI	(PRE, CON:sub)	
unterordnende Konjunktion mit Satz	weil, daß, damit, wenn, ob	KOUS	CON:sub	Konjunktion-subordinierend
nebenordnende Konjunktion	und, oder, aber	KON	CON:coo	Konjunktion-koordinierend
Vergleichskonjunktion	als, wie	KOKOM	(CON:sub)	
"zu" vor Infinitiv	zu [gehen]	PTKZU	?	
abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad	PTKVZ	–	
Antwortpartikel	ja, nein, danke, bitte	PTKANT	?	
Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]	PTKA	–	
Komma	,	\$,	PON:comma	Satzzeichen-Komma
Satzbeendende Interpunktion	. ? ! ; :	\$.	PON:sep	Satzzeichen-!-.-...-:-;-?
sonstige Satzzeichen; satzintern	—[,]()	\$(	PON	Satzzeichen-sonstige
Kompositions-Erstglied	An- [und Abreise]	TRUNC	–	
		(tag depends on abreviated word)	ABR	Abkuerzung
Nichtwort, Sonderzeichen enthaltend	D2XW3	XY	SYM	Symbol

4. Tag Set Characteristics for Better Comparability

In the previous chapter, we have explored the various factors which may lead to non-equivalence in the PoS-annotation of the subcorpora of a multilingual corpus. Of all of the factors we mentioned, the first one is the easiest to manipulate: appropriate design of the tag set can, indeed, help make the information we extract from each subcorpus more comparable. Thus we should ask: what are the characteristics of tags and sets of tags that contribute to equivalence of the PoS-annotation in each subcorpus?

4.1. Medium Granularity

First of all, it is generally better for tag sets to make too many distinctions than too little: apart from the obvious fact that finer distinctions allow the extraction of more precise information, we have seen in section 3.1.3 that splitting categories is safe, whereas merging them can lead to difficulties because information is lost in the process.

At the same time however, the high tagging accuracy which is also a prerequisite for comparability is difficult to obtain with a very detailed tag set, because it requires distributional information for a greater number of tags to be available, which can only be learnt from a larger training corpus (see section 2.1).

Considering this opposition more closely reveals that it is simply another internal/external validity—or computational tractability/linguistic soundness—trade-off such as the one we encountered in section 3.1.3, which shows how sometimes the tag set designers' desire for tagging accuracy can lead to tag set simplifications.

4.2. Hierarchicality

Another point is that in order to help neutralise granularity problems, it is desirable that the labels chosen as tags be based on the ten traditional major part-of-speech categories, possibly subdivided into subcategories, and that they be organised into a hierarchical, tree-like structure, meaning that all labels under a common mother should share a common stem (Leech, 1997).

The reason for this is that it allows to use wildcards in queries if the querying language supports them: underspecified queries at different levels of granularity can be formulated when one tag set is more fine-grained than the other with respect to a given PoS-category.

Fortunately, modern tag sets almost always have such a logical structure: for example, in the Stuttgart-Tübingen tag set (Schiller et al., 1995), which differentiates between coordinating conjunctions, subordinating conjunctions introducing finite and infinitive

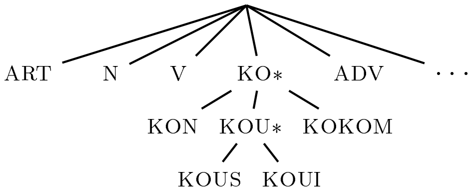


Figure 2: A partial, simplified tree representation of the
Stuttgart-Tübingen tag set, showing the subtree for conjunctions

clauses, and conjunctions introducing the complement of a comparative, all tags for conjunctions (German "**Ko**njunktion") start with **KO** and those for subordinating (German "**U**nterordnend") conjunctions with **KOU**, thus allowing the extraction of all conjunctions, or of all subordinating conjunctions, by using respectively `KO*` or `KOU*`[9] in the query (see figure 2).

Figure 3 illustrates what happens when this is not the case: in the pronoun-subtree of the Stuttgart-Tübingen tag set, there is actually no way to retrieve all relative pronouns, because <PWAV> does not share a enough with <PRELS> and <PRELAT>—a problem which would remain even if the unfortunate collapsing of adverbial interrogative and adverbial relative pronouns under a single tag (see section 3.1.3) had not already disrupted the correspondence with Stein (1995)'s <PRO:rela>.

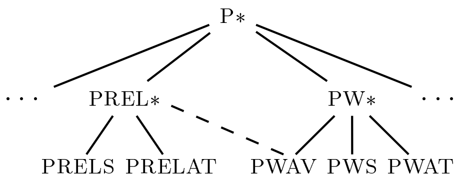


Figure 3: The non-tree-like structure of the tag labels
in the pronoun-subtree of the STTS

4.3. Guideline Compliance

Finally, we also mentioned in section 3.1.4 that divergences among linguistic theories can be a source of non-equivalence. This problem has in part been solved by proposals made in recent years by two international efforts, the EAGLES initiative (Leech and Wilson, 1996) and the MULTEXT project (Ide and Véronis, 1994, Calzolari and Monachini, 1996), which have published recommendations on the design of tag sets in order to help harmonise tagging practises.

One important point in these recommendations is that, recognising that there will remain differences in the annotation due to differences between languages and between the needs of specific applications, they especially stress what they called the 'principle of extensibility', which allow the introduction of categories relevant only to one language, or of categories particularly useful for the end-use of the corpus to be annotated.

[9] Again, as in section 3.2.1, the * here is the Kleene star, a wildcard operator which replaces any combination of characters.

5. Dealing with Non-Equivalence in Information Extraction

Finally, we will take a look at what can be done when we attempt to extract information from a multilingual corpus, but are faced with incomparable PoS-annotation in the subcorpora.

Fortunately, in many cases, there are ways we may be able to help ourselves out, of which we will see a few examples in section 5.2, after we describe *Le Monde Diplomatique* French-German parallel corpus, from which these examples are drawn.

5.1. The *Le Monde Diplomatique* Parallel Corpus

The texts composing the *Le Monde Diplomatique* French-German parallel corpus were taken from the CD-Rom archive of the French monthly newspaper *Le Monde Diplomatique* (2001), of which a German edition is also available, published by *taz, die tageszeitung*.[10] Only those articles which were accompanied by a German translation on the CD-Rom were selected.

In its present state, the corpus comprises the relevant articles from the January, February, March, October, November and December 1999 issues, altogether 136 articles in each language. It contains almost 243 000 word tokens for French and a little over 224 000 for German; for information, some corpus statistics are displayed in figure 4, which were computed with WordSmith Tools.[11]

	Source files	Target files
Tokens	242 695	224 230
Types	20 700	27 747
Type/Token Ratio	8.53	12.37
Standardised Type/Token	48.78	54.55
Sentences	9 385	9 218
Sentence length	25.86	24.32
Standardised Sentence Length	15.57	13.24

Figure 4: Some corpus statistics

[10] See `http://www.taz.de`.

[11] WordSmith Tools Version 3.0 (for Windows), developed by Mike Scott, copyright Oxford University Press, 2002. Demo version available for download from `http://www.oup.com/elt/global/isbn/6890/`

5.1.1. Building the Corpus

In order to be able to extract parallel syntactic information from the corpus, the three major steps in building the corpus were

1. its 'alignment', a process by which a correspondence is established between sentences or segments in both versions of the texts;
2. its tagging with the TreeTagger's French and German parameter files (Schmid, 1994, 1995, Stein and Schmid, 1995);[12]
3. and its preprocessing for use with Gsearch, a software tool designed to extract from corpora sentences matching specific syntactic criteria, and which runs under Unix-like operating systems (Corley et al., 2001a,b).[13]

Aligning

In order to increase tagging accuracy, before starting to align, the files were spell-checked and the data was somewhat "cleaned up". Two Perl scripts were used for this, one for French and one for German, which processed the raw text in the following way:

- Words beginning a paragraph which were printed in capital letters were lowercased, except for the initial letter, so that they could be found in the tagger's lexicon;

- Non-continuous text was removed: copyright indications, bibliographical references, page numbers, information on type of text (such as *fiche documentaire*, etc.), authors' names and qualifications (e. g. *directeur de. . . , historien, écrivain, auteur de. . .*) footnotes, keywords, text id-numbers, etc.;

- References to the footnotes were also removed.

Pairs of French-German files were then aligned with the alignment tool contained it the software Déjà Vu.[14] The alignment units were in general full sentences, i. e. main clauses containing a finite verb together with their dependent subordinate clauses. Material which was missing in one of both versions of the text was deleted, and in some cases where the splitting or joining of sentences in the two versions of the texts made aligning too difficult, the material was simply removed.

In addition, some small corrections were made to the text, such as rectifying typos, adding spaces before or after dashes to differentiate them from hyphens (thus avoiding the tagging of all words followed by a dash as <TRUNC> in the Stuttgart-Tübingen tag set), and deleting remaining references to footnotes.

Following this, in order to be able to tag each subcorpus with the respective parameter files, it was necessary to re-separate the aligned files into two subcorpora containing the

[12] The TreeTagger and its Linux and Unix parameter files for English, German, French and Italian are freely available for download from `http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html`

[13] Available for download from `http://www.hcrc.ed.ac.uk/gsearch/`

[14] Déjà Vu is a translation memory programme for Windows aimed at professional translators, of which a 30-day evaluation version is available for download from `http://www.atril.com`. For the alignment of the corpus, we used the file alignment wizard included in the database maintenance utilities.

French and German texts respectively, by exporting the result of the alignment of a file pair into two "source" (French) and "target" (German) files in raw text format, in which the segments were delimited by quotation marks.

Tagging

In the next step, a Perl script converted these "source" and "target" files to Unix file format (by calling the Unix command-line tool `dos2unix --ascii`), enclosed the segments in SGML tags replacing the quotation marks delimiters, and also uniquely numbered each segment in order to keep track of the correspondence between segments across languages.

The files were then tagged with the TreeTagger and its parameter files for French and German, which use the Stuttgart-Tübingen (Schiller et al., 1995) and Stein (Stein, 1995) tag sets. Unfortunately, the scripts provided along with the TreeTagger to simplify the tagging process contained some hard-coded options, and thus, as such, were of no use to us. However, in order to work straight from the Linux command line, a French tokeniser and a French filter to correct common tagging mistakes had to be created, by splitting up the script code provided with the tagger.

The files were then tokenised, taking proper care of abbreviations, then tagged with the respective parameters for each language,[15] and finally the respective correction filters were applied.

Setting up Gsearch

Having thus tagged each subcorpus separately, the data was put back together in order to provide parallel concordances: the parallel corpus was re-built by a Perl script which printed alternatively one source and one target segment, so that each French segment would be followed by the corresponding German segment in the final corpus. At the same time, an additional column was added to the final corpus, which indicates the language of each segment.

Finally, the corpus had to be be translated, or 'filtered', into the format required internally by Gsearch. Thus a filter was written (`gsearch_ttag.pl`, see appendix p. 32), the corpus processed with it, and Gsearch was setup (see the file `.gsearchrc` in the appendix on p. 33) so that it can find the filtered corpus and knows what fields it contains, namely tokens, parts-of-speech, lemmata and language, thus now allowing queries on the basis of all four types of fields in the final version of the corpus.

5.1.2. A Parallel Corpus, but not a Translation Corpus

At this point, it should be mentioned that the resulting corpus should be thought of as a parallel corpus, but that it cannot really be considered as a translation corpus.

[15] When calling the tagger, we used options to print the tokens, tags, and lemmata, as well as options to skip the SGML tags and to use the aligned segments as tagging units, instead of relying on sentence punctuation. The command line was:
`$tree-tagger <language-parameters> -token -lemma -sgml -eos-tag seg`

La loi sur la sécurité nationale, dans son article 7, prévoit des peines de prison ferme pour toute activité "anti-étatique", *ce qui inclut les prises de position favorables au régime de Pyongyang ou à l'idéologie communiste - "une violation flagrante des libertés fondamentales", souligne M. Kwak Nohyun.*	Das Gesetz zur nationalen Sicherheit sieht in Artikel 7 für jede "gegen den Staat" gerichtete Handlung Gefängnisstrafen ohne Bewährung vor.
⋮	⋮
	Nach einer weiteren Phase höchster Spannungen zwischen den beiden Ländern hat Washington Mitte September bekanntgegeben, dass das seit sechsundvierzig Jahren andauernde Wirtschaftsembargo gegen Nordkorea teilweise aufgehoben werden soll. Zuvor hatte sich das Regime in Piöngjang bereit erklärt, seine Tests mit Langstreckenraketen auszusetzen. Von einer Entspannung ist allerdings im Lande selbst nicht viel zu spüren; dort kämpfen die Menschen nach wie vor ums nackte Überleben, während das Regime "Normalitt" simuliert und sich mit einem dicken Wall aus Propanganda und Misstrauen umgibt. Aber auch Südkorea tut sich schwer mit dem Ende des Kalten Krieges.
De Séoul, plusieurs agences de tourisme offrent la possibilité de jeter un coup d'oeil sur le pays ennemi - la Corée du Nord - en organisant des visites de la DMZ, la zone démilitarisée qui sépare les deux Corées.	In Seoul organisieren mehrere Reiseveranstalter Ausflüge in die entmilitarisierte Zone (DMZ), die Nord- und Südkorea voneinander trennt und von der aus man einen Blick auf das feindliche Nachbarland werfen kann.
⋮	⋮
15 août 1948. La proclamation de la République de Corée (au sud) met officiellement fin à l'occupation américaine. Syngman Rhee est élu président.	15. August 1948: Die Proklamation der Republik Korea (im Süden) beendet offiziell die US-amerikanische Besatzung, *obgleich 35 000 amerikanische Soldaten auf der Halbinsel verbleiben.* Syngman Rhee wird zum ersten Präsidenten gewählt.
9 septembre 1948. Proclamation de la République populaire démocratique de Corée (RPDC), dont Kim Il-sung devient président.	9. September 1948: Proklamation der Demokratischen Volksrepublik Korea (DVRK). Präsident wird Kim Il Sung, *der dieses Amt bis zu seinem Tod im Jahre 1994 ausübt.*

Figure 5: Missing segments in both the French and German versions of a November 1999 article series on North and South Korea (the missing parts are printed in italics)

L'arrivée récente d'indigènes awas sur la terre qu'ils ont démarquée a provoqué des conflits (*six* maisons brûlées notamment).	Als Awa-Indianer ein Stück Land für sich absteckten, gingen *sieben* Häuser in Flammen auf.
En 1992, le rapport du groupe de travail de l'AP insista en ce sens et, en juin 1995, un autre rapport du groupe de travail prit position en faveur de la création des unités de visites familiales.	In einem Bericht von *Juni 1995* sprach sich auch eine Arbeitsgruppe der Abteilung Strafvollzug in diesem Sinne aus; in einem weiteren Bericht setzte sie sich für die Schaffung von Langzeitbesuchsräumen ein.

Figure 6: Two excerpts from February 1999 articles, which contain different figures and dates (in italics) in the French and German versions

Although the corpus was originally built for the purpose of studying how particular French words or structures are rendered in their German translations, in the process of aligning the texts extracted from the CD-Rom, it became obvious that we cannot properly speak of "originals" and "translations" (or of "source texts" and "target texts") when referring respectively to the *Le Monde Diplomatique* and the *taz* versions of the texts, because the articles appear more like parallel versions of the same stories than like translations of stories from one language into another.

Indeed, there are often whole sentences, sometimes even whole paragraphs, which are present in only one of both languages, be it French or German, as can be seen in the excerpt shown in figure 5—in the case of a translation, it would be understandable that the German text have been simplified, but not that the original texts contain less information than its translation—and that there are also quite often differences in the figures, dates, etc. in both versions of an article (see figure 6).[16]

Consequently, the only reasonable assumption seems to be that both articles were created on the basis of a common draft, the language of which is unclear (possibly sometimes French, sometimes German, but also maybe other languages, since there are also English and Spanish editions of *Le Monde Diplomatique*), and that the final authors thoroughly rewrote their versions, sometimes changing the figures to be printed, thus resulting in two parallel versions of the same text, among which it cannot be said that one is the "original" and the other the "translation".

5.1.3. Querying the Corpus

In the corpus, each French segment is followed by the corresponding German segment, so that parallel concordances can be obtained by using the Gsearch query options `-b1` and `-a1`, which require that Gsearch output the previous or the next sentence in the corpus, for German and French queries respectively. Moreover, sequential queries can also be done by using the options `-sc` ("subcorpus") for the output and `-sf` ("subcorpus file") for the input, and running the script `transform-subcorpus.pl` (see appendix p. 34) in between the queries, which circumvents the fact that Gsearch treats output context together with original sentences as one sentence in subsequent queries (see sample queries in figure 7).

The information which the TreeTagger outputs being of three kinds—words, tags (called `pos` in Gsearch queries) and lemmata—all three can be used as query objects, in addition to which the language of the resulting matches (`fr` or `de`) can also be specified. The Gsearch query language also allows the user to use regular expressions,[17] and supports various options, which are described in Corley et al. (2001a,b).

Moreover, additional power and flexibility in the queries can be obtained by using two very simple phrase-structure grammars[18] (PSGs) for French and German, which

[16] These observations were made by myself and by another student, Véronique Nessius, who helped me extract the texts from the *Le Monde Diplomatique* CD-Rom and align the texts of the corpus.

[17] At first glance, regular expressions are somewhat comparable to wildcards in other query languages. A "`.`" (dot) stands for "any character", and various operators are used to specify how many times the preceding character may appear in the result: `a*`, for instance, means the letter `a` may appear any number of times (including 0). See any good introduction to regular expressions, for example in Church (1994).

General usage: gsearch [<options>] <corpus> <grammar> <goal>
(in the case of the *Le Monde Diplomatique* parallel corpus, the corpus name is `concord` and the grammars are `GrammarFR` and `GrammarDE` for French and German respectively)
Simple queries:
Building a parallel concordance for the French words ending in *-ent*:

```
$ gsearch -a1 [further options] concord - "<word=.*ent & language=fr>"
```

Building a parallel concordance for the adjectives derived from the German word *zahl*:

```
$ gsearch -b1 [further options] concord - "<word=zahl.* & pos=ADJA>"
```

Querying for the sentences in which a noun phrase (as defined in the PSG) is followed by *est*:

```
$ gsearch [further options] concord GrammarFR np "<word=est>"
```

Queries requiring sequential Gsearch calls:
Looking for the cases in which the French word *présenter* was translated by *präsentieren* in German:

```
$ gsearch -a1 [further options] -sc concord - "<lemma=présenter>" > [TMP1]
$ transform-subcorpus.pl [TMP1] > [TMP2]
$ gsearch -b1 [further options] -sf [TMP2] concord - "<lemma=präsentieren>"
```

Figure 7: The syntax used to query the corpus with Gsearch.

are adaptations to the Stuttgart-Tübingen and Stein tag sets of the simple grammars distributed with Gsearch (`GrammarDE` is partially illustrated in figure 8; for a full listing, see appendix p. 37). However, in their present state, these grammars do not offer a realistic coverage of any syntactic phenomenon, so that in reality, they would probably first have to be extended depending on particular extraction needs.

5.2. Some Ad-hoc Extraction Methods

Using the corpus described in the previous section, we can now take a concrete look at what can be done in some cases when we want to extract syntactic information from a multilingual corpus but are faced with non-equivalent PoS-annotation.

Depending on what information it is we want to extract, there are various possible solutions, among which we will examine three cases in the following.

5.2.1. Extraction Based on Specific Word Forms

First of all, when we cannot do a query on the basis of the part-of-speech information, there is always the possibility of doing a query on the basis of specific word-forms (such as endings, etc.), which is actually what we first attempted to do in the example in the introduction.

For example, remember we mentioned in section 3.1 that both French and German have present participles, but that the Stuttgart-Tübingen tag set does not have a tag for present participles. Knowing that present participles are used very differently in French and German, imagine we wanted to compare their uses in both languages—we would then

[18] A typical phrase-structure grammar is made up of a set of rules of the form `A → B C`, which specify that the syntactic constituent on the left is composed of a sequence of other elements, which may be either shorter constituents or basic elements (most often words, but also tags in the case of a tagged corpus). Thus in a toy PSG-grammar, we could for example have rules like `S → NP VP`, meaning that a sentence may be composed of a noun phrase (the subject) followed by a verb phrase, but also rules like `NP → the N`, which means that noun phrases can be made up of the article *the* followed by any noun, and lexical rules like `N → cat` and `VP → sleeps`.

```
vp --> verb
vp --> verb np                      adj  --> <ADJA>
vp --> verb pp                      prep --> <APPR.*>
                                    det  --> <ART>
np --> det n1+                      noun --> <NN>
np --> n1+                          det  --> <P.*T>
n1 --> adj* noun+                   verb --> <V.FIN>
                                    verb --> <VVIMP>
pp --> prep np
```

Figure 8: The German phrase-structure grammar for use with Gsearch (excerpt)

need to extract the present participles from the German subcorpus without resorting to PoS-information.

In order to do this, we could put to use the fact that German present participles are formed by adding *-end* to the verb stem (*sprech+end*, Engl. “to speak”, *tu+end*, Engl. “to do”), except for verbs ending in *-ern* and *-eln*, which drop the *e* (*feier+nd*, Engl. “to celebrate”, *segel+nd*, Engl. “to sail”). Moreover, since German present participles are often used as adjectives, we must allow for inflection (comparison degrees and declension), so that the Gsearch query would become:[19]

```
$ gsearch -M concord GrammarDE "<word=.+el*r*nd(er)*(st)*e*[rsnm]* & language=de>"
```

This query finds 1159 sentences in our corpus. However, when examining the result, we see that the precision[20] achieved in the result is quite low: in a 10 % sample (116 sentences) randomly extracted with Gextract (a tool distributed with Gsearch), we have 83 sentences which indeed contain a present participle and 33 which do not (but instead contain such words as *während*, Engl. “during/while”, a preposition or conjunction, various compounds of *Tausend*, Engl. “thousand”, the verb *wenden*, Engl. “to turn”, etc.)

Since the tagger also outputs the lemma for each word, one possibility of refining our query would be to add as a condition that the lemma of the matching words must end in *-end*:

```
$ gsearch -M concord GrammarDE "<word=... & language=de & lemma=.+el*r*nd>"
```

However, this does not help much: of the 174 sentences which are thus eliminated (so that only 985 remain), 89 are rightly rejected, but in the 85 remaining cases, adding the condition on lemmata leads to rejecting correct sentences (false negatives).

Thus we see that extraction on the basis of word forms does not seem to be a reliable means to make up for lacking PoS-annotation: apart from the additional effort required to formulate the query in this way, we have seen that in the example we have just discussed, the results are not very good. Moreover, it must be added to this that in many cases, there is no way to recover with an acceptable accuracy the information we want: such is for instance the case with the French participles example we mentioned in the introduction.

[19] In this command, the Gsearch option `-M` instructs the parser to search for multiple matches of the goal in a sentence or segment.

[20] The only measure we can compute to evaluate most our queries is precision; except for the first query in section 5.2.3, we cannot estimate the recall rate for our queries because the number of false negatives is unknown.

5.2.2. Extraction Using Parallel Concordances

Another way we might be able to recover the information we want to extract would be to put to use the fact that we are working with a multilingual corpus: if the tag set for one subcorpus does not allow the retrieval of a given structure, it might be possible to retrieve it indirectly by building a parallel concordance of the corresponding structures in another language.

Concretely, we will examine two such cases in our test corpus, both of which involve tags which exist in the Stuttgart-Tübingen tag set but not in Stein's tag set (see section 3.1).

The first example concerns the tagging of the French words *de*, *à*, *pour*, etc., which Stein's tag set, in accordance with the French linguistic tradition, tags as a prepositions even when they introduce a subordinate infinitive clause—as opposed to other linguistic traditions, which often distinguish between the cases where the corresponding words occur before nouns and those where they introduce an infinitive clause (Engl. *He asked me to<TO> help him*, Susanne and Penn-Treebank tag sets, Garside et al. (1987), Santorini (1991);[21] German *Er bat mich, ihm zu<PTKZU> helfen*, STTS, Schiller et al., 1995).

Since the Stuttgart-Tübingen tag set has an extra category for the German *zu* in this context, it might thus be possible to extract French infinitive clauses introduced by a preposition by building a parallel concordance for the particle *zu* in German, using the following Gsearch query, which returns 1060 sentences:[22]

```
$ gsearch -M -p pos -b1 concord - "<pos=PTKZU>"
```

Unfortunately however, in this case at least, this approach does not return the expected results: in a manual examination of 53 sentences (a 5 % random sample) containing <PTKZU>, only 24 were found in which <PTKZU>+infinitive corresponded to *de*/*à*/etc. followed by an infinitive in French. The reason for this seems to be that either different verbs were used in French and German to render the same idea, or that the subcategorisation frames of the French and German verbs in corresponding versions of a text are too often different.

Our second example is hardly more encouraging. Another word category which Stein's tag set does not distinguish are the prepositions, adverbs and conjunctions which introduce a comparison. We attempted to retrieve them from the corpus by building a parallel concordance of the German sentences containing comparisons. The query we used for this is:

```
$ gsearch -M -p pos -b1 concord - "<pos=KOKOM>"
```

In answer to this query, Gsearch returned 978 sentences from our test corpus, of which we selected a 10 % random sample (98 sentences) for manual inspection.

Of these, 59 were not taken into account because the tagger had incorrectly attributed the tag <KOKOM> to other uses of the German words *als* and *wie*, such as sentences in which a verb subcategorises for a prepositional phrase introduced by *als* (*gelten als*),

[21] See also the TnT homepage (`http://www.coli.uni-sb.de/~thorsten/tnt/`), which uses both tag sets.

[22] `-p pos`: show the PoS tags for the search goal.

temporal uses of *als* meaning “when”, the locution *ebenso wie* which is synonym of *and*, etc.

Among the remaining 39 however, only 29 sentences containing a <KOKOM> in German corresponded to a comparison construction in French, whereas 10 (25.6 %) did not—thus confirming the results of the first experiment.

5.2.3. Extraction Requiring Parsing

Finally, when we are not able to extract a given structure from a corpus strictly on the basis of the part-of-speech annotation, other syntactic information may allow to recover the desired structures, due to the redundancy inherent to natural language.

For example, we have observed in section 3.1 that the Stuttgart-Tübingen tag set has a single tag for adverbial relative pronouns and adverbial interrogative pronouns, <PWAV>. In this case, word order could help identifying at least direct interrogative pronouns, which usually occur at the beginning of a sentence, and after which subject and verb are inverted, the verb occurring before its subject, as in *Wo<PWAV> wohnt<VVFIN> er<PPER>?* (Engl. “Where does he live?”).

Using this fact, a straightforward query for partial direct interrogative sentences (i. e. those introduced by a wh-word) would be the following, for which we only added to the grammar illustrated in figure8 that pronouns (<PDS>, <PIS>, <PPOSS>, <PWS> and <PPER>) are also nominal phrases:

```
$ gsearch concord GrammarDE "<pos=PWAV>" "<pos=V.FIN>" np
```

This query gives back 14 sentences, but it actually misses some sentences, as can be shown by running in sequence the two following queries, which use punctuation to retrieve all direct interrogative sentences and then, from that, extract those containing a <PWAV>, finally capturing 19 sentences:[23]

```
$ gsearch -sc -P concord GrammarDE "<word=\? & language=de>" > interro_sc.txt
$ gsearch -sf interro_sc.txt concord GrammarDE "<pos=PWAV & language=de>"
```

By examining the difference between both files, we observe that our first query was too simple: of course, in some cases, there may be words such as adverbs or reflexive pronouns between the finite verb and the nominal phrase, so that the original query should have been:

```
$ gsearch concord GrammarDE "<pos=PWAV>" "<pos=V.FIN>" "<pos=ADV>*" "<pos=PRF>*" np
```

Unfortunately, this still returns only 16 matches, and misses sentences (15.8 %) in which anything more than a single adverb or a reflexive pronoun occurs between verb and subject, as well as sentences in which the subject nominal phrase is even only slightly more complex than that allowed presently in our grammar, such as:

[23] `-P`: take punctuation into account (which otherwise is ignored by the Gsearch parser).

```
<s1>Und
<pos PWAV>wo <pos VVINF>bleiben <pos APPR>bei <pos ADV>alledem <pos ART>die
<pos NN>Länder
des Südens?</s1>

<s1>
<pos PWAV>Warum <pos VVFIN>tut <pos PRELS>die <pos ADV>einst <pos
ADJD>multilateral <pos ADJA>gestimmte <pos NN>Clinton-Administration
alles, um die UNO zu diskreditieren?</s1>
```

These sentences cannot be said to be very complex, yet in order to capture them, we would have to refine our grammar in more than one way.

In the first case, we would have to add that short prepositional phrases can also occur between verb and subject, thus requiring us to decide what a "short" prepositional phrase is, and we would also probably have to hard-code *bei alledem* as a prepositional phrase, since it is a somewhat irregular form and the tagger wrongly considers *alledem* to be an adverb instead of a pronoun.

As for the second case, it would require writing a much more complex grammar for German nominal phrases, which would allow adjective phrases modifying a noun to occur between this noun and its determiner.

This short example illustrates the fact that even for fairly simple cases—we did not even attempt to distinguish indirect adverbial interrogative pronouns from relative pronouns, something which would require determining the attachment site, as we saw in section 3.1.3—extraction by parsing will inevitably require gradual extension of the grammar and refining of the query, a process which often turns out to be tedious and may lead to very large grammars, such as Zamparelli's and Lapata's two "large scale" grammars for English noun phrases (164 lines) and verb phrases (103 lines) of the British National Corpus, which are distributed along with Gsearch.

Thus we see that, although it is theoretically possible to retrieve what we want by parsing, this requires more effort than when we can extract it on the basis of the PoS-annotation: first of all, a fairly powerful query language which allows parsing is needed (fortunately provided by Gsearch's use of phrase-structure grammars) and in addition, the necessary grammars must be created, something which is not a trivial task.

6. Conclusion

In this paper, we took a look at why the part-of-speech annotation in subcorpora of a multilingual corpus is often incomparable, at how this problem could in part be remedied, and at what can be done to retrieve syntactic information from the corpus in spite of non-equivalence in the PoS-annotation.

After having reviewed the linguistic foundations of automatic PoS-tagging, in order to help us understand how different factors can influence the tagging process, we detailed the following sources of non-equivalence:

- particular syntactic properties of each language;
- differing granularity of tag sets with respect to some PoS-categories;
- interaction between linguistic criteria and computational tractability;
- vagueness of linguistic theories, or divergence of theoretical stances;
- inclusion of semantic information in some tags;
- tokenisation difficulties for some words (multiwords and mergers);
- tagging "in context" or independently from it for some PoS-categories;
- tagger accuracy and systematic tagging errors.

Following this, we asked ourselves what characteristics of a tag set tend to make the part-of-speech annotation of the subcorpora of a multilingual corpus more comparable: if a tag set were to be developed or adapted with multilingual information extraction in mind, to what aspects should attention be paid in particular?

Our answer to this question was that tag sets used in multilingual information extraction should, if possible, be of medium granularity, be hierarchical, and comply to international guidelines such as those proposed by the EAGLES initiative and by the MULTEXT project (Leech and Wilson, 1996, Ide and Véronis, 1994, Calzolari and Monachini, 1996).

Finally, we explored what can be done when one is faced with non-equivalence of the PoS-annotation in the subcorpora of a multilingual corpus. We saw that if the information contained in the part-of-speech annotation is not sufficient to allow the extraction of a given syntactic structure, there are various other ways we may still be able to retrieve the syntactic structures we want from the corpus.

In order to illustrate this, we first described the French-German *Le Monde Diplomatique* parallel corpus, which we built for use as a test corpus. We then considered three different ways of retrieving syntactic structures from the corpus when the PoS-annotation is of no help: extraction based on specific word-forms, extraction using parallel concordances, and extraction requiring parsing.

However, the examples we gave showed that in most cases when we are forced to find alternatives because we cannot rely on the PoS-annotation, the precision of the result is much lower. Moreover, it also became obvious that finding an ad-hoc solution for each structure that cannot be extracted straightforwardly often turns out to require a lot of extra work: each problem must be examined separately and the query refined by trial-and-error, to which the additional effort of grammar development must be added in in the case of parsing-based extraction.

Appendix: Scripts Used to Build the *Le Monde Diplomatique* Parallel Corpus

In the following, we will briefly comment on some of the more interesting scripts which were written to help build and query the *Le Monde Diplomatique* parallel corpus.

1. Filtering the Corpus for Use with Gsearch

In order to be able to use Gsearch on the corpus, it first had to be translated into Gsearch's internal format, which is described in section 6.4 of the Gsearch manual. This was done by a Perl filter, `gsearch_ttag.pl`, which was written by adapting the `raw.pl` filter provided with Gsearch.

```
#! /usr/bin/perl

###############################################################################################
# Description: Gsearch filter that takes French or German text tagged with the H. Schmid's
# TreeTagger (IMS, Stuttgart) and outputs it in the internal format required by Gsearch
###############################################################################################

# Command line arguments: -c<corpusname> -F<inputfiles> (see section 6.4 of the Gsearch manual)
# <corpusname> is a comment which may be used to name the corpus
# <inputfiles> are the files to be filtered

($corpusname, $inputfiles) = ("@ARGV" =~ /\-c (.*) \-F (.*)/);

# Get date and time
require "ctime.pl";
$date = &ctime(time);

# Print a header to keep gsearch happy
print "\% CORSET corpus file\n";
print "\% Created from $corpusname\n";
print "\% Processed using the filter gsearch_ttag.pl\n";
print "\% Created on $date";
print "\% Corpus files: $inputfiles\n\n";

@files = split(/\s+/, $inputfiles);
# Split $inputfiles, using spaces as delimiter
# @files is an array containing all the words in $inputfiles

foreach $file(@files) {
    print "\n\-$file\n"; # print file name

    open(INFILE, $file) || die "Can't open file $file\n";

    # Initialize a flag to remember whether the preceding line ended in an apostrophe
    $apostrophe=0;

    while ($line = <INFILE>) {
        if (($line !~ /<\/seg>/) &&
```

```
            ($line !~ /<seg_id=([0-9]{1,4})/)) {
            # Punctuation signs should be preceded by *
            # A. Stein's punctuation tags are PON PON:comma PON:sep
            # STTS punctuation signs are \$, \$. \$(
            if (($line =~ /\$/) || ($line =~ /PON/)) {
                print "*";
                # Commas and end-of-sentence punctuation (i.e. punctuation that has no space
                # before it) should preceded by +
                if (# comma
                    ($line =~ /\$,/) || ($line =~ /PON:comma/) ||
                    # punctuation between sentences
                    ($line =~ /\$\./) || ($line =~ /PON:sep/))
                { print "+"; }
            }
            # If the preceding line ended in an apostrophe,
            # the line should be preceded by a +
            elsif ($apostrophe==1) { print "+"; $apostrophe=0 }
            # If this line ends in an apostrophe, memorize that
            elsif ($line =~ /\'\W/) { $apostrophe=1 }

            print "\t$line";
        }

        # Instead of relying on the punctuation to determine sentences, we rely on the SGML
        # tags we introduced as delimiters. These segments should separated by 2 or more \n
        # (but since each $line ends in a trailing \n, we only add one \n)
        if ($line =~ /<\/seg>/) { print "\n"; }
    }
    close(INFILE);
}
```

2. Setting up Gsearch

In addition to the central `Setup` file, Gsearch also allows each user to have a `.gsearchrc` file in his home directory, in which the location of specific corpora, the filters to be used and the fields contained in each corpus are defined.

The difficulty here was that the exact syntax for the `Files` field was not documented in the Gsearch manual: although the manual says that the `Files` field is a regular expression, it does not mention that the slashes indicating directories should be written as `(/` instead of `/`, that there should be as many `)?`'s at the end of the line as there are `(/`'s, not does it say that dots should be escaped (`\.`).

Moreover, the manual does not mention either that no comments are allowed within the corpus specification fields.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% File: ~/.gsearchrc
% Description: Corpus definition file for using Gsearch on the Le Monde Diplomatique corpus
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

word-tag-lemma-lg {
  Fields: word pos lemma language
}

concord : word-tag-lemma-lg {
        CorpusName: Le Monde Diplomatique - Tagged Concordance
        Filtered:   .../Corpora/MondeDiplo/Concord/filtered_concord.gsearch
        Files:      1999(/..(/.*\.cc)?)?
        Filter:     empty
    }
```

3. Circumventing Gsearch's Treatment of Context

Gsearch's default behaviour is to treat matched sentences and context output by the `-a` or `-b` options as one sentence for subsequent queries. However, since we make use of these options to output parallel concordances, this behaviour was not desirable in the case of our corpus. We thus wrote a script which, when run between subsequent calls to Gsearch, uses the `language` field in our corpus to separate the matched sentences from the output "context".

Two examples illustrate this: in the first, four sequential queries were run, the last of which with a German search criterion and the `-b1` option to output the corresponding French segments, and in the second, a French criterion was given and German context output with option `-a`. In both cases, a excerpt from the resulting subcorpus is shown first, followed by the modified subcorpus after our script was run.

As the examples show, in the original version of the subcorpus, there is no newline and no sentence number between the French and German segments, but this has been rectified by the script in the modified version of the subcorpus.

First example: Original Subcorpus

```
% CORSET corpus file
% Created using gsearch
% Basic args: gsearch concord - <language=de>
% Created on 01/19/03 15:07:11
% Corpus files: SUBCORPUS

-1999/01/01_siecle.cc
>               (original sentence=s2)
>               [orig](original sentence=s2)
>               [orig][orig](original sentence=s2)
>               [orig][orig][orig](original sentence=s2)
                Quelles         PRO:REL         quel            fr
                sont            VER:pres        être            fr
(snip)
                la              DET:ART         le              fr
                planète         NOM             planète         fr
*+              ?               PON:sep         ?               fr
                An              APPR            an              de
                der             ART             d               de
(snip)
                zu              PTKZU           zu              de
                machen          VVINF           machen          de
*+              .               $.              .               de

>               (original sentence=s3)
>               [orig](original sentence=s4)
>               [orig][orig](original sentence=s4)
>               [orig][orig][orig](original sentence=s16)
                Les             DET:ART         le              fr
                Etats-Unis      NAM             Etats-Unis      fr
(snip)
                aucune          PRO:IND         aucun           fr
                difficulté      NOM             difficulté      fr
*+              .               PON:sep         .               fr
                Denn            KON             denn            de
                auch            ADV             auch            de
(snip)
```

First example: Modified Subcorpus

```
% CORSET corpus file
% Created using gsearch
% Basic args: gsearch concord - <language=de>
% Created on 01/19/03 15:07:11
% Corpus files: SUBCORPUS

-1999/01/01_siecle.cc
>                (original sentence=s1)
                 Quelles         PRO:REL          quel            fr
                 sont            VER:pres         être            fr
(snip)
                 la              DET:ART          le              fr
                 planète         NOM              planète         fr
*+               ?               PON:sep          ?               fr

>                (original sentence=s2)
                 An              APPR             an              de
                 der             ART              d               de
(snip)
                 zu              PTKZU            zu              de
                 machen          VVINF            machen          de
*+               .               $.               .               de

>                (original sentence=s15)
                 Les             DET:ART          le              fr
                 Etats-Unis      NAM              Etats-Unis      fr
(snip)
                 aucune          PRO:IND          aucun           fr
                 difficulté      NOM              difficulté      fr
*+               .               PON:sep          .               fr

>                (original sentence=s16)
                 Denn            KON              denn            de
                 auch            ADV              auch            de
(snip)
```

Second example: Original Subcorpus

```
-1999/01/01_siecle.cc
>                (original sentence=s3)
                 Les             DET:ART          le              fr
                 Etats-Unis      NAM              Etats-Unis      fr
(snip)
                 jamais          ADV              jamais          fr
                 fait            VER:pper         faire           fr
*+               .               PON:sep          .               fr
                 Geopolitisch    ADJD             <unknown>       de
                 beherrschen     VVFIN            beherrschen     de
(snip)
                 Land            NN               Land            de
                 zuvor           ADV              zuvor           de
*+               .               $.               .               de
```

Second example: Modified Subcorpus

```
-1999/01/01_siecle.cc
>                (original sentence=s3)
                 Les             DET:ART          le              fr
                 Etats-Unis      NAM              Etats-Unis      fr
(snip)
                 jamais          ADV              jamais          fr
                 fait            VER:pper         faire           fr
*+               .               PON:sep          .               fr
```

```
>               (original sentence=s4)
                Geopolitisch     ADJD            <unknown>       de
                beherrschen      VVFIN           beherrschen     de
(snip)
                Land             NN              Land            de
                zuvor            ADV             zuvor           de
*+              .                $.              .               de
```

The script `transform_subcorpus.pl`

```
#!/usr/bin/perl

#############################################################################################
# Description: Use between sequential Gsearch queries on the Le Monde Diplomatique corpus,
# if the 1st query used option -a or -b (circumvents Gsearch default behaviour)
#############################################################################################

open(SC,"$ARGV[0]") || die "Could not open subcorpus\n";

while (defined ($line = <SC>)) {
    if (# Header, newlines and "-1999..." should be left alone
        ($line =~ /^%/ ) || ($line eq "\n" ) || ($line =~ /^-/ ))
    {   # Chomp off \n between filename and sentence number ("-1999...")
        if ($line =~ /^-/ ) { chomp($line); }
        # Headers and such are not part of the French or German subcorpora
        $language = "none";
        print $line;
    } elsif (# In sequential queries, Gsearch adds an ">(original...)" line each time on top of
             # the previous ones, and adds an "[orig]" prefix before the old ones, e.g.:
             # -1999/01/01_siecle.cc
             # >       (original sentence=s2)
             # >       [orig](original sentence=s2)
             # >       [orig][orig](original sentence=s2)
             # >       [orig][orig][orig](original sentence=s2)
             # We want to remember only the sentence number of the line with the highest number
             # of "[orig]" prefixes (loop)
             $line =~ /^>\t(\[orig\]){0,}\(original sentence=s([0-9]{0,})/ ) {
        $language = "none";
        $sentencenumber = $2;
    } elsif (# Circumvent default Gsearch treatment of context by introducing a blank line
             # followed by an ">(original...)" line between French and German segments
             $line =~ /\t.*\t.*\t.*\t(..)\n/ ) {
        $language = $1;

        # When query criteria is German, with French context being output before it, the sen-
        # tence number after the filename is even and belongs to the German sentence; it must
        # be first decreased to obtain French sentence number, printed, and then re-increased
        if ($language ne $prev_lang) {
            if (($language eq "fr") && (($sentencenumber % 2) == 0))
            { --$sentencenumber; }
            $orig = "\n>\t\(original sentence=s" . $sentencenumber . "\)\n" ;
            print $orig;
            ++$sentencenumber;
        }
        print $line;
    }
    $prev_lang=$language;
}

close SC;
```

4. The Grammars for Use with Gsearch

Finally, we list here for reference the simple phrase-structure grammars which we adapted to the Stein (1995) and Stuttgart-Tübingen (Schiller et al., 1995) tag sets. At present these grammars are not much more than toy grammars, but they build the starting point for developing more interesting grammars.

The Grammar for French, Adapted to the Stein (1995) Tag Set

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Description: Simple grammar distributed with Gsearch, adapted to the A. Stein's tag set
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

#defterm "pos" % Set default terminal field: <pos=ADJA> can now be written as <ADJA>

%%%%%%%%%%%%%%%
% Grammar rules
%%%%%%%%%%%%%%%

vp --> v1
vp --> v1 np
vp --> v1 pp
v1 --> adv* verb adv*
np --> det n1+ pp*
np --> n1+ pp*
np --> np conj np
n1 --> adj* noun+
n1 --> n1 conj n1
pp --> prep np

%%%%%%%%%%%
% Terminals
%%%%%%%%%%%

%    --> <ABR>             % Abkuerzung
adj  --> <ADJ>             % Adjektiv
adv  --> <ADV>             % Adverb
%    --> <ADV:neg>         % Adverb-Negation
conj --> <CON:coo>         % Konjunktion-koordinierend
%    --> <CON:sub>         % Konjunktion-subordinierend
det  --> <DET:def>         % Artikel-definit
det  --> <DET:indef>       % Artikel-indefinit
det  --> <DET:part>        % Artikel-partitiv
%    --> <INT>             % Interjektion
noun --> <NOM>             % Nomen
noun --> <NPR>             % Eigenname
%    --> <NUM>             % Zahl/Zahladjektiv
prep --> <PRE>             % Praeposition
prep --> <PRE:det>         % Praeposition-mit-Artikel-verschmolzen
%    --> <PRO:clit>        % Pronomen-klitisch-y/en
%    --> <PRO:demo>        % Pronomen-demonstrativ
%    --> <PRO:indef>       % Pronomen-indefinit
%    --> <PRO:inter>       % Pronomen-interrogativ
%    --> <PRO:pers:clit:1> % Pronomen-personal-klitisch-1.Person
%    --> <PRO:pers:clit:2> % Pronomen-personal-klitisch-2.Person
%    --> <PRO:pers:clit:3> % Pronomen-personal-klitisch-3.Person
%    --> <PRO:pers:conj:1> % Pronomen-personal-unbetont-1.Person
%    --> <PRO:pers:conj:2> % Pronomen-personal-unbetont-2.Person
%    --> <PRO:pers:conj:3> % Pronomen-personal-unbetont-3.Person
%    --> <PRO:pers:disj:1> % Pronomen-personal-betont-1.Person
%    --> <PRO:pers:disj:2> % Pronomen-personal-betont-2.Person
%    --> <PRO:pers:disj:3> % Pronomen-personal-betont-3.Person
%    --> <PRO:pers:reci:1> % Pronomen-personal-reziprok-1.Person
%    --> <PRO:pers:reci:2> % Pronomen-personal-reziprok-2.Person
```

```
%    --> <PRO:pers:reci:3> % Pronomen-personal-reziprok-3.Person
%    --> <PRO:poss>        % Pronomen-possessiv
%    --> <PRO:rela>        % Pronomen-relativ
verb --> <VER:cond.*>      % Verb-conditionnel
verb --> <VER:futu.*>      % Verb-futur
verb --> <VER:impf.*>      % Verb-imparfait
%    --> <VER:infi>        % Verb-Infinitiv
%    --> <VER:pper>        % Verb-Partizip-Perfekt
%    --> <VER:ppre>        % Verb-Partizip-Praesens
verb --> <VER:pres.*>      % Verb-present
verb --> <VER:simp.*>      % Verb-passesimple
verb --> <VER:sub.*>       % Verb-subjonctif
```

The Grammar for German, Adapted to the STTS Tag Set

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Description: Simple grammar distributed with Gsearch, adapted to the Stuttgart-Tübingen tag
% set and slightly extended
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

vp --> v1
vp --> v1 np
vp --> v1 pp
v1 --> adv* verb adv*
np --> pn
np --> det n1+ pp*
np --> det n1+
np --> n1+ pp*
np --> n1+
np --> np conj np
np --> pron
n1 --> adj* noun+
n1 --> n1 conj n1
pp --> prep np
rc --> rpro np+ v1

%%%%%%%%%%%
% Terminals
%%%%%%%%%%%

adj  --> <ADJA>     % attributives Adjektiv
%    --> <ADJD>     % adverbiales oder prädikatives Adjektiv
adv  --> <ADV>      % Adverb
prep --> <APPR>     % Präposition; Zirkumposition links
prep --> <APPRART>  % Präposition mit Artikel
%    --> <APPO>     % Postposition
%    --> <APZR>     % Zirkumposition rechts
det  --> <ART>      % bestimmter oder unbestimmter Artikel
%    --> <CARD>     % Kardinalzahl
%    --> <FM>       % Fremdsprachliches Material
%    --> <ITJ>      % Interjektion
%    --> <ORD>      % Ordinalzahl
%    --> <KOUI>     % unterordnende Konjunktion mit ‘‘zu’’ und Infinitiv
%    --> <KOUS>     % unterordnende Konjunktion mit Satz
conj --> <KON>      % nebenordnende Konjunktion
%    --> <KOKOM>    % Vergleichskonjunktion
noun --> <NN>       % normales Nomen
pn   --> <NE>       % Eigennamen
pron --> <PDS>      % substituierendes Demonstrativpronomen
pron --> <PIS>      % substituierendes Indefinitpronomen
pron --> <PPOSS>    % substituierendes Possessivpronomen
pron --> <PWS>      % substituierendes Interrogativpronomen
det  --> <PDAT>     % attribuierendes Demonstrativpronomen
det  --> <PIAT>     % attribuierendes Indefinitpronomen ohne Determiner
det  --> <PPOSAT>   % attribuierendes Possessivpronomen
det  --> <PWAT>     % attribuierendes Interrogativpronomen
```

```
%    --> <PIDAT>    % attribuierendes Indefinitpronomen mit Determiner
pron --> <PPER>     % irreflexives Personalpronomen
rpro --> <PRELS>    % substituierendes Relativpronomen
rpro --> <PRELAT>   % attribuierendes Relativpronomen
%    --> <PRF>      % reflexives Personalpronomen
%    --> <PWAV>     % adverbiales Interrogativ- oder Relativpronomen
%    --> <PAV>      % Pronominaladverb
%    --> <PTKZU>    % ``zu'' vor Infinitiv
%    --> <PTKNEG>   % Negationspartikel
%    --> <PTKVZ>    % abgetrennter Verbzusatz
%    --> <PTKANT>   % Antwortpartikel
%    --> <PTKA>     % Partikel bei Adjektiv oder Adverb
%    --> <TRUNC>    % Kompositions-Erstglied
verb --> <VVFIN>    % finites Verb, voll
verb --> <VVIMP>    % Imperativ, voll
%    --> <VVINF>    % Infinitiv, voll
%    --> <VVIZU>    % Infinitiv mit ``zu'', voll
%    --> <VVPP>     % Partizip Perfekt, voll
verb --> <VAFIN>    % finites Verb, aux
%    --> <VAIMP>    % Imperativ, aux
%    --> <VAINF>    % Infinitiv, aux
%    --> <VAPP>     % Partizip Perfekt, aux
verb --> <VMFIN>    % finites Verb, modal
%    --> <VMINF>    % Infinitiv, modal
%    --> <VMPP>     % Partizip Perfekt, modal
```

Bibliography

Abeillé, A. and Clément, L. (1999). Désambiguation morpho-syntaxique. Guide pour les annotateurs—mots composés. Corpus Le Monde, version 0.4. Until 1st February, 2002 available from `http://talana.linguist.jussieu.fr/~lionel/corpus/guide-mc.ps`.

Abeillé, A., Clément, L., and Kinyon, A. (2001). Building a treebank for French. In Abeillé, A., editor, *Treebanks* (forthcoming). Kluwer Academic Publishers, The Netherlands. Until 1st February, 2002 available from `http://talana.linguist.jussieu.fr/~lionel/corpus/kluwer.ps`.

Calzolari, N. and Monachini, M. (1996). MULTEXT. Common specifications and notation for lexicon encoding. Available from `http://www.lpl.univ-aix.fr/projets/multext/LEX/LEX1.html`.

Church, K. W. (1994). Unix for poets. Available from `http://www.research.att.con/wc/tutorials/unix_for_poets.ps` (html and pdf versions also available at various places on the web).

Corley, M., Corley, S., Crocker, M. W., Keller, F., and Trewin, S. (2001a). *Gsearch User Manual*. Revision 1.3 of the manual for Gsearch version 2.06w. Available from `http://www.hcrc.ed.ac.uk/gsearch/src/manual.ps.gz`.

Corley, S., Corley, M., Keller, F., Crocker, M. W., and Trewin, S. (2001b). Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

Déjean, H. (2000). How to evaluate and compare tagsets? A proposal. In *Proceedings of LREC2000*, Athens, Greece. Available from `http://lcg-www.uia.ac.be/lcg/ps/dejean.lrec2000.ps.gz`.

Garside, R., Leech, G., and Sampson, G. (1987). *The Computational Analysis of English. A Corpus-Based Approach,* appendix B: The Susanne tag set. Longman, London. Cited by Déjean (2000). See also the TnT homepage (`http://www.coli.uni-sb.de/~thorsten/tnt/`), which uses the Susanne tag set.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation. Linguistic Information from Computer Text Corpora*, chapter 7. Longman, London.

Grévisse, M. and Goosse, A. (1989). *Nouvelle Grammaire Française.* Duculot, Louvain-la-Neuve, Belgium, 2nd edition.

Ide, N. and Véronis, J. (1994). MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 588–592, Kyoto, Japan. COLING'94. Available from `http://www.up.univ-mrs.fr/~veronis/pdf/1994multext.pdf`.

Le Monde Diplomatique (2001). Archives 1980-2000. 21 années du Monde Diplomatique en texte intégral. CD-Rom. Order from `http://www.monde-diplomatique.fr/cederom/`.

Leech, G. (1997). Grammatical tagging. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation. Linguistic Information from Computer Text Corpora*, chapter 2. Longman, London.

Leech, G. and Wilson, A. (1996). Recommendations for the morphosyntactic annotation of corpora. Expert Advisory Group on Language Engineering Standards (EAGLES). Available from `http://www.ilc.pi.cnr.it/EAGLES96/browse.html`.

Manning, C. and Schütze, H. (1999). *Foundations of statistical and natural language processing*, chapter 10. MIT Press, Cambridge, MA.

Rajman, M., Lecomte, J., and Paroubek, P. (1997). Format de description lexicale pour le français, partie 2: Description morpho-syntaxique. Available from `http://www.limsi.fr/TLP/grace/doc/GTR-3-2.1.ps.gz`.

Sampson, G. (1995). *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford.

Santorini, B. (1991). Part-of-speech tagging guidelines for the Penn treebank project. Available from `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.ps`.

Schiller, A., Teufel, S., and Stöckert, C. (1995). Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/stts_guide.ps.gz`.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK. Revised version of the paper available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.ps.gz`.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop*. Revised version of the paper available from `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.ps.gz`.

Stein, A. (1995). Liste der Tags in französischen Korpora. Available from `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`.

Stein, A. and Schmid, H. (1995). Étiquetage morphologique de textes français avec un arbre de décisions (Morphological tagging of French texts using a decision tree). *t.a.l. (Traitement Automatique des Langues)*, 36 (Traitement probabilistes et corpus—Probabilistic Approaches to Corpora)(1–2):23–36. Available from `http://www.uni-stuttgart.de/lingrom/stein/pubs/steins95.ps.gz`.

Teufel, S., Schmid, H., Heid, U., and Schiller, A. (1996). Study of the relation beteen tagsets and taggers. Expert Advisory Group on Language Engineering Standards (EAGLES). Available from `http://www.ilc.pi.cnr.it/EAGLES96/browse.html`.

Véronis, J. and Khouri, L. (1995). Étiquetage grammatical multilingue: le projet Multext. *t.a.l. (Traitement Automatique des Langues)*, 36(1–2):233–248. Draft available from the Multext project website: `http://www.lpl.univ-aix.fr/projets/multext/LEX/LEX2.html`.

Witten, I. H. and Frank, E. (2000). *Data Mining*. Morgan Kaufmann.