Towards an integrated representation of multiple layers of linguistic annotation in multilingual corpora

Elke Teich & Silvia Hansen
Institute for Applied Linguistics, Translation and Interpreting (FR 4.6)
University of Saarland, Germany
E.Teich@mx.uni-saarland.de / S.Hansen@mx.uni-saarland.de

1. Introduction: Annotation of natural language corpora

There has been an increasing interest in recent years in the enrichment of natural language corpora in terms of annotation with explicit linguistic information. This interest manifests itself most prominently in two areas of linguistics: corpus linguistics and computational linguistics. For corpus linguistics, the long standing practice has been to work on raw, i.e., unannotated text. While raw corpora are basically fine for some kinds of linguistic work, notably for lexicology and lexicography, for other kinds of linguistic analysis tasks, e.g., for syntactic or semantic analysis, the information that needs to be extracted is not readily derivable from raw text. Thus, corpora have to be annotated with linguistic categories in order to be able to extract the desired kinds of information. For such annotation to be practicable at all, the annotation process needs to be carried out automatically or at least semi-automatically.

The automatic processing of large corpora, including linguistic annotation, has been a central issue in computational linguistics in recent years. Here, one of the main interests is in the statistical processing of natural language data (cf. Charniak 1993), such as statistically-based part-of-speech tagging or statistical parsing. The main purpose of these techniques is application in natural language systems, such as, for instance, in machine translation (e.g., Brown et al 1990). These techniques can also be employed for purposes of corpus-based, descriptive linguistics. In recent years, most of the large corpora of English (BNC, LOB, Bank of English etc.) have been annotated with part of speech information, which has made it possible to exploit them also for syntactic analysis. Also corpora with shallow syntactic annotation (annotation at phrase structure level) exist (e.g., the Penn Treebank (Marcus et al 1993)).

What remains problematic, however, is linguistic annotation at more abstract levels of linguistic organization, notably the semantic and discourse strata. Here, annotation can only be carried out semi-automatically, e.g., with the help of tools that support interactive mark-up of texts by humans. If a corpus is to be annotated with more than one kind of annotation, we find ourselves in a situation in which the corpus exists in a number of versions, one for each kind of annotation, e.g., a syntactic one and a semantic one. This has some serious implications for the exploitation of the corpus for information extraction in that it is impossible to query the corpus with reference to more than one layer of annotation at a time. This problem has been increasingly acknowledged both in corpus linguistics and in computational linguistics. One of the paradigms proposed to overcome such difficulties is the one of document encoding, a paradigm that has been increasingly applied in humanities computing, including linguistic applications (e.g., TEI (Sperberg-McQueen & Burnard 1999), XCES (XCES 2000).

The present paper is concerned with the issue of the integration of different kinds of linguistic annotation for multilingual corpora employing the paradigm of document encoding using the Extensible Mark-up Language (XML). The context in which this is of interest for us is corpus-based translation analysis; more specifically, what we are interested in is the empirical testing of hypotheses concerning the specific properties of translations when compared to original texts in the same languages as the target language and to original texts in the source language. The paper is organized as follows. First, we briefly present our analysis scenario (Section 2). Then we discuss the annotation techniques we have employed to enrich our corpora with the desired linguistic information (Section 3). In Section 4 we present a possible solution to the integration of different kinds of corpus annotation. Section 5 concludes the paper with a summary and discussion of issues for future work.

2. Empirical investigations of the specific properties of translations

The assumption that translations have particular properties that distinguish them from non-translations, i.e., original texts, has been around for many years. For instance, some researchers in translatology have observed that translations are more explicit than the original texts they are translated from; or, that translations use simpler language than originals, e.g., their vocabulary being less varied than that of comparable texts in the same language as the target language (TL) (cf. Baker 1995, 1996, Kenny 2001, Sager 1994, Steiner 2001, Thome 1975, Toury 1995). However, until recently, such hypotheses have been merely suspicions that were only supported by rather anecdotal evidence. With the increasing availability of natural language corpora, it begins to be possible to empirically test such hypotheses, using parallel corpora (source language (SL) texts and their translations) and comparable corpora (translations and TL original texts).

One example of such an analysis scenario is the Translational English Corpus (TEC) (Baker 1995, 1996), for which the British National Corpus (BNC) acts as the comparable TL original corpus. The following hypotheses about the specific, and possibly universal, properties of translations have been proposed (Baker 1995, Laviosa-Braithwaite 1996):

Explicitation. Translations show a tendency to spell things out rather than leave them implicit. An indicator for explicitation is text length, translations tending to be longer than monolingually comparable original texts. Also, some language-specific tests have been proposed, e.g., for English, frequency counts of optional elements, such as 'that' as complementizer and as relative pronoun, have been suggested, translations tending to use such elements more frequently than comparable original texts.

Simplification. Translations tend to use simpler language than original texts in the same language as the TL, possibly to optimize the readability of the target language text. Possible measures for simplification are average sentence length, lexical density and type-token ratio, the latter being a standard measure for the vocabulary variation in a text.

Normalization. Translations have a tendency to conform to the typical patterns of the TL, exaggerating the typical features of the TL. As a test for normalization, Baker (1996) suggests comparing the use of punctuation (translations purportedly using punctuation less creatively than comparable texts in the same language as the TL).

Levelling out. In a collection of translations compared to a collection of comparable original texts in the same language as the TL, the individual texts in the set of translations are more similar to each other than the individual texts in the set of original texts. For levelling out, some of the above mentioned measures can be applied; one would then predict that for translations, the extreme values for lexical density, type-token ratio and average sentence length are closer to each other than for original texts.

The measures suggested to test these hypotheses (text length, sentence length, lexical density, type-token ratio) are rather shallow linguistic features that essentially operate on words. This kind of information can be straightforwardly extracted from a raw corpus using some standard functions provided by concordance tools such as WordSmith (Scott 1996). However, measures such as text length or type-token ratio are only of limited value, if we want to arrive at an interpretation of such a kind of analysis. For instance, if it can be shown that translations are longer than comparable texts in the TL (or SL originals), there is still the question of why this would be the case. One possible source of explanation lies in information packaging: a text in which information is more densely packed is shorter than a text in which information is less densely packed (e.g., more complex nominal groups + simple clause structure vs. more complex clause structure + simple nominal groups). In order to get at syntactic information such as clause structure or the structure of phrases, however, a corpus needs to be annotated with syntactic categories.

In our own work (Hansen 1999, Hansen & Teich 2001, Teich & Hansen 2001, Teich 2001), we propose a number of tests for the above introduced hypotheses that are more theoretically-informed. However, carrying out such tests requires to enrich the corpus with explicit linguistic information, such as part-of-speech (PoS) tags, shallow syntactic structure, semantic features as well as discourse features. Another example of more abstract linguistic information being needed in order to arrive at linguistically-meaningful interpretations of the hypotheses formulated is normalization: Here, what is needed is a well-defined notion of what "normal" means in language. One candidate notion of "normality" is the notion of register, i.e., of functional linguistic variation (cf. e.g., Biber 1990, Halliday 1985, Quirk et al 1985). Registers are typically described as sets of texts that exhibit significant frequencies of co-occurring grammatical features. The level at which these features are described varies across approaches: Some operate at a rather shallow syntactic level (e.g., in Biber's work),

others are functional and thus more abstract (e.g., in Halliday's work). If we want to use the notion of register to test for normalization in translations, again, we need to annotate the corpus under investigation in terms of the relevant linguistic categories. A third example of the kinds of information one may be interested in a corpus-based analysis of translations are textual features, notably cohesion. This is particularly interesting in the domain of simultaneous interpreting, a situation of linguistic online-production under severe time pressure. One of the assumptions in relation to the hypothesis of explicitation is that simultaneous interpretations differ from their SL originals in the deployment of cohesive means, e.g., tending to avoid pronominal reference and employing lexical cohesion instead (Kusztor, personal communication). Again, in order to extract information about the cohesive patterns prevalent in a corpus, the corpus needs to be annotated with the relevant kinds of linguistic information.

What is thus called for is the annotation of the corpus in terms of various layers or strata of linguistic organization. As will be seen in the following section, annotations at different layers can be carried out automatically or at least semi-automatically, using a range of tools such as part-of-speech taggers, syntactic parsers and support tools for manual annotation for more abstract linguistic features. While the application of these techniques is straightforward, there are some remaining problems, however (see Section 3).

3. Linguistic annotation and information extraction from annotated corpora

The most basic techniques for corpus-based analysis (word counts, word lists, KWIC concordances) are available in most standard concordance tools, such as e.g., WordSmith (Scott 1996). These basic techniques typically operate on raw text and linguistic annotation is not necessary.

As explained in the previous section, if more abstract linguistic information is to be extracted, texts have to be annotated with the relevant information first, before any analysis can be carried out. Depending on how abstract the linguistic features to be analyzed are, the linguistic corpus annotation can be done automatically or it must be done semi-automatically, i.e., by means of manual annotation with computer support.

In the following we describe a number of techniques that range from automatic part-of-speech tagging, shallow parsing and translation alignment to semi-automatic techniques of annotation.

Part-of-speech tagging and **shallow parsing**. Part-of-speech tagging is carried out fully automatically, either using a rule-based or a statistical approach, where recently, statistical approaches prevail. For multilingual applications, it is important that the tagger can be used for more than one language. Automatically analyzing a corpus in terms of syntactic structure is still a challenging task and cannot be carried out with satisfactory accuracy yet. To avoid this problem, recently researchers in computational linguistics who are interested in the accurate parsing of large amounts of text promote what has been called *interactive parsing*, where a parser carries out a shallow parse and a human may correct or add information to the proposed parse. For example, the parser assigns syntactic phrase labels to the elements of a clause, but does not resolve syntactic ambiguities of particular kinds, such as PP-attachment, leaving this to the human to deal with. Also, the parser may assign syntactic function labels to constituents, but these will be checked by a human.
One system which combines part-of-speech tagging and shallow parsing is the ANNOTATE system (Plaehn & Brants 2000) under development in the TIGER and NEGRA projects (Brants 2000a). ANNOTATE uses the TnT tagger (Brants 2000b) that can be applied multilingually and has been trained on a number of languages, including English and German. The tag set used for English is the Susanne tag set (Sampson 1995); the one for German is based on the Stuttgart-Tübingen tag set (Hinrichs et al 1995). TnT includes a tool for tokenization, which is a preparatory step in the tagging process. Furthermore ANNOTATE uses Cascaded Markov Models (CMM (Brants 1999a, 1999b)) for the analysis of phrase categories as well as grammatical functions. For this reason terminal nodes (for parts-of-speech and morphology), non-terminal nodes (for phrase categories) and edges (for grammatical functions) are labeled during the interactive annotation with ANNOTATE (see Figure 1).

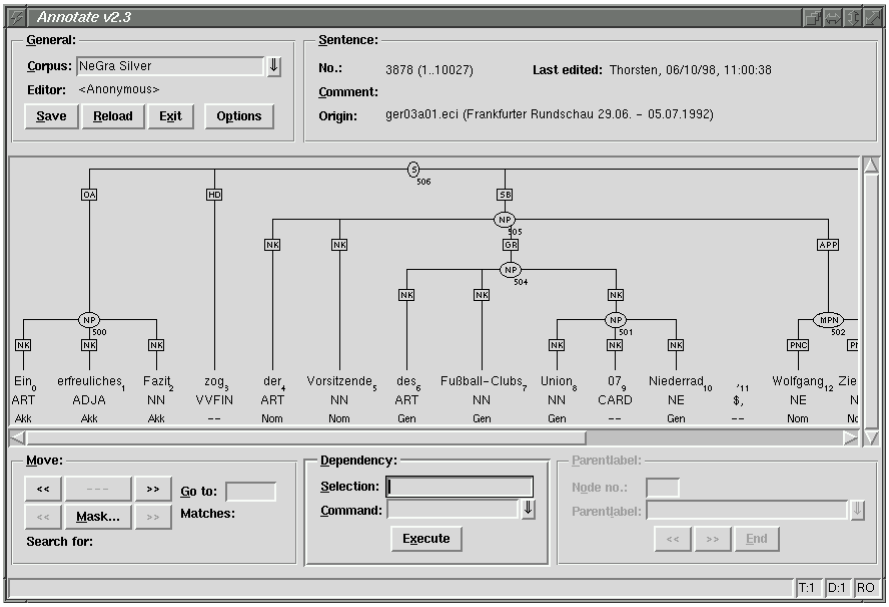

Figure 1: Interactive annotation with ANNOTATE[1]

The tagged and parsed corpus data are stored in the form of a relational database, but can be exported to text format. Figure 2 shows an example of the representation of corpus annotation produced by ANNOTATE.

```
#BOS 1 15 892541360 1
Mögen           VMFIN   3.Pl.Pres.Konj          HD      508
Puristen        NN      Masc.Nom.Pl.*           NK      505
aller           PIDAT   *.Gen.Pl        NK      500
Musikbereiche   NN      Masc.Gen.Pl.*           NK      500
auch            ADV     --                      MO      508
die             ART     Def.Fem.Akk.Sg  NK      501
Nase            NN      Fem.Akk.Sg.*            NK      501
rümpfen         VVINF   --                      HD      506
,               $,      --                      --      0
die             ART     Def.Fem.Nom.Sg          NK      507
Zukunft         NN      Fem.Nom.Sg.*            NK      507
der             ART     Def.Fem.Gen.Sg          NK      502
Musik           NN      Fem.Gen.Sg.*            NK      502
liegt           VVFIN   3.Sg.Pres.Ind           HD      509
für             APPR    Akk                     AC      503
viele           PIDAT   *.Akk.Pl                NK      503
junge           ADJA    Pos.*.Akk.Pl.St         NK      503
Komponisten     NN      Masc.Akk.Pl.*           NK      503
im              APPRART Dat.Masc                AC      504
Crossover-Stil  NN      Masc.Dat.Sg.*           NK      504
.               $.      --                      --      0
#500            NP      --                      GR      505
#501            NP      --                      OA      506
#502            NP      --                      GR      507
#503            PP      --                      MO      509
#504            PP      --                      MO      509
#505            NP      --                      SB      508
#506            VP      --                      OC      508
#507            NP      --                      SB      509
#508            S       --                      MO      509
#509            S       --                      --      0
#EOS 1
```

Figure 2: Sample annotation of ANNOTATE in text format[2]

[1] The screenshot of ANNOTATE is taken from the project web page: http://www.coli.uni-sb.de/sfb378/negra-corpus/screenshot.html

The first column in Figure 2 displays the word or the referring parent node, the second column the part-of-speech tag, the third one the morphological analysis, the fourth one the edge and the last one the parent. Figure 3 shows the same annotation in Penn Treebank format.

```
(S
        (S-MO
                (VMFIN-HD Mögen)
                (NP-SB
                        (NN-NK Puristen)
                        (NP-GR
                                (PIDAT-NK aller)
                                (NN-NK Musikbereiche)
                        )
                )
                (ADV-MO auch)
                (VP-OC
                        (NP-OA
                                (ART-NK die)
                                (NN-NK Nase)
                        )
                        (VVINF-HD rümpfen)
                )
        )
        ($, ,)
        (NP-SB
                (ART-NK die)
                (NN-NK Zukunft)
                (NP-GR
                        (ART-NK der)
                        (NN-NK Musik)
                )
        )
        (VVFIN-HD liegt)
        (PP-MO
                (APPR-AC für)
                (PIDAT-NK viele)
                (ADJA-NK junge)
                (NN-NK Komponisten)
        )
        (PP-MO
                (APPRART-AC im)
                (NN-NK Crossover-Stil)
        )
    ($. .)
)
```

Figure 3: Sample annotation of ANNOTATE in Penn Treebank format[3]

For extraction of text instances tagged and parsed with ANNOTATE, information retrieval tools like the IMS Corpus Workbench (Christ 1994) can be employed. This system allows queries for words and/or annotation tags on the basis of regular expressions. An important feature of the system for multilingual application is that its Corpus Query Processor (CQP) caters for querying of parallel corpora (i.e., source language texts and translations). For an example of a query executed on a parallel English-German corpus see Figure 4.

[2] The sample annotation of ANNOTATE is taken from the project web page: http://www.coli.uni-sb.de/sfb378/negra-corpus/corpus-sample.export

[3] The sample annotation of ANNOTATE in Penn Treebank format is taken from the project web page: http://www.coli.uni-sb.de/sfb378/negra-corpus/corpus-sample.penn

```
# Query: DE_EN; passives-de = [pos="VB.*"] [] {0,1} [pos="VVN.*"];
#------------------------------------------------------------------------------------------------------------------------
     149:    , and how good it was to  <be driven> . He tilted his face to c
-->de_de:    Wie gut , daß die Nacht vorüber war , wie gut , jetzt gefahren zu werden .
     729:    newspaper . A ferry had  <been sunk> just off the island . ' I
-->de_de:    In den Gewässern vor der Insel war eine Fähre gesunken .
     850:    country ' s future will  <be decided> today . Yours too , perha
-->de_de:    Zukunft des Landes entscheidet sich heute .
     927:    nced , because shots had  <been fired> at a remote polling stati
-->de_de:    Der Schriftsteller und er müßten aufbrechen , in einem abgelegenen Wahllokal
             seien Schüsse gefallen .
     943:    tion that one person had  <been killed> and another badly wounded
-->de_de:    Er verschwieg einen Toten und eine Schwerverletzte ; jede Hilfe für sie käme zu
             spät , hatte ihm Romulus leise gemeldet .
    1037:    d how firmly the buttons  <were sewn> on . Two women lifted Kur
-->de_de:    sie zeigten Schnitt und Größe , wiesen auf Markennamen hin und führten die
             Festigkeit der Knöpfe vor .
```

Figure 4: Sample query with CQP

Semantic annotation. When more abstract linguistic features are to be coded, annotation must be carried out manually. Tools supporting manual annotation typically carry out some basic operations automatically, such as segmentation of a text into units of annotation, and provide a few functions for interpretation (e.g., basic descriptive statistics, querying of the annotated corpus). The central functions of such tools, however, are that they provide the possibility of defining annotation schemes that are tailored to the specific needs of the user and that they support the annotation of a corpus using a defined scheme. One such tool is Coder (O'Donnell 1995). For an example of Coder's interface for annotation scheme definition see Figure 5. The example scheme essentially says that a clause can carry the semantic feature 'agentive' or is not marked for agentivity ('non-agentive') or that the category does not apply at all ('not-applying').

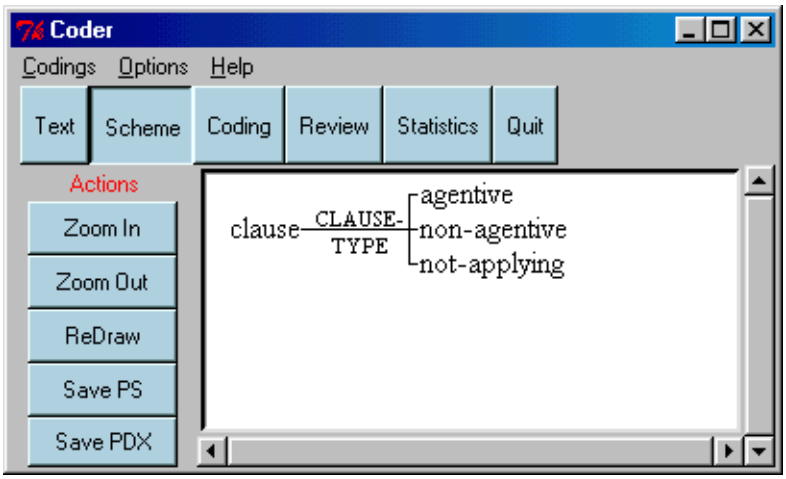

Figure 5: Coder's interface for annotation scheme definition

Texts annotated with Coder are represented in an XML/SGML-like format, as shown in Figure 6.

```
<segment features="clause non-agentive" comment="" ignore=0>
        15  Thus in water, H2O, the hydrogen atoms exchange between
        different oxygen atoms billions of times per second.
</segment>
<segment features="clause not-applying" comment="" ignore=0>
        16  In some compounds, namely acids, the molecules are so averse
        to the hydrogen they contain
</segment>
<segment features="clause non-agentive" comment="" ignore=0>
        that they will readily donate the hydrogen to other molecules.
</segment>
<segment features="clause not-applying" comment="" ignore=0>
        17  One such is hydrogen chloride, HCl,
</segment>
<segment features="clause non-agentive" comment="" ignore=0>
        and textbooks often write this process as HCl    H+ + Cl-.
</segment>
```

Figure 6: Coder output format

Alignment. For the analysis of a parallel corpus, the units of translation (source language text units and their translations) need to be aligned. There are various alignment programs freely available, for example the Déjà Vu program (Atril 2000). Déjà Vu aligns a text and its translation sentence by sentence, storing the aligned texts in one file, where the source language sentence and its translation are represented in a tab separated vector (TSV) format. Files of aligned texts created by Déjà Vu can be exported to translation workbenches and to some standard applications, such as MS Excel or MS Access. Figure 7 gives an example of a piece of aligned text (English originals and German translations).

```
"16 In some compounds, namely acids, the molecules are so averse to the hydrogen
they contain that they will readily donate the hydrogen to other molecules."
        "16  In einigen Verbindungen - besonders in Saeuren - sind die
Abstossungskraefte zwischen Molekuel und Wasserstoff so gross, dass der
Wasserstoff sofort an andere Molekuele abgegeben wird."

"17 One such is hydrogen chloride, HCl, and textbooks often write this process as
HCl H+ + Cl-."        "17a  Salzsaeure,  HCl,  ist  eine  solche
Verbindung. 17b In Lehrbuechern wird dieser Prozess oft durch die Gleichung HCl H+
+ Cl- dargestellt."
```

Figure 7: Déjà Vu aligned text

While it is generally possible to annotate a corpus in terms of various layers of linguistic information (shallow-syntactic, functional-grammatical, semantic, discoursal) making use of the techniques just described, there is one remaining problem: A given corpus will exist in various versions, one for each kind of annotation in the worst case, and it can thus only be queried with respect to one kind of annotation at a time. If, however, we want to make reference to more than one kind of annotation, we need an integrated representation of the annotated corpus, where each kind of annotation is represented as one tier that stands in a well-defined relation to the other tiers. In our analysis scenario, for instance, we may want to test the hypothesis that English employs more non-agentive NP Subjects than German (cf. Doherty 1993), and if this is true, we would like to see whether in German translations from English there is interference concerning this feature or whether there is compensation. When annotations such as ‘non-agentive’, ‘NP’, ‘Subject’ reside in separate encodings of the corpus, it is not possible to extract instances of non-agentive NP Subjects in a straightforward way.

At a more general level, the problem we encounter here is the problem of integration of heterogeneous sources, a well known problem in the context of data base integration. Since this problem has typically not merely to do with format transformations, but involves more fundamental questions of data representation, we will suggest that one possible way towards a solution is to apply the paradigm of document encoding using the Extensible Mark-up Language (XML).

4. Towards an integrated representation of multiple layers of corpus annotation

Each of the tools used for encoding, annotation and extraction employs different input formats that do not necessarily match straightforwardly. For instance, the IMS Corpus Workbench requires as input a tokenized text with syntactic annotations in a TSV format, and Coder requires as input raw text (segmentation is done within Coder). Also, the outputs that are generated for the different kinds of annotation are again different across tools: ANNOTATE produces a TSV format, Coder produces an XML/SGML-like format.
In order to be able to query the corpus making reference to more than one kind of annotation at the same time, the different kinds of annotation have to be merged into one uniform representation (see Figure 8).

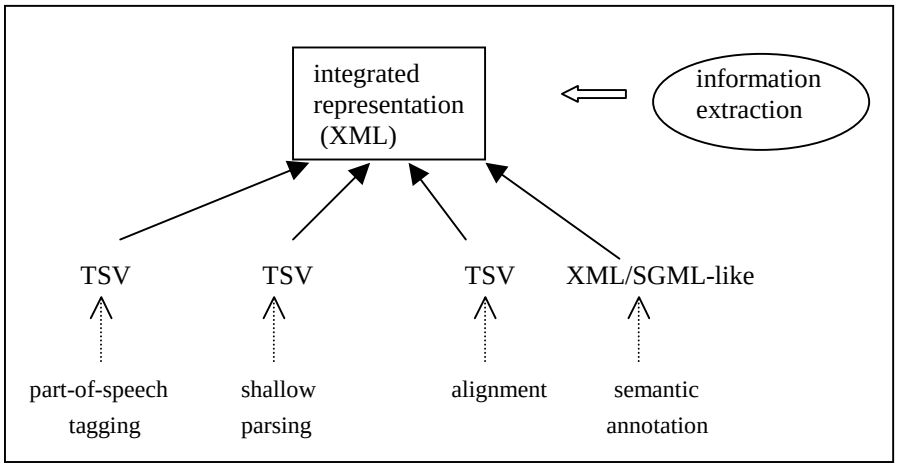

Figure 8: Integrated representation

This merging can be carried out straightforwardly by using Perl scripts and XSLT style sheets (W3C-XSLT 2000), but there are some more principled questions involved here to do with the fact that we do multi-layer annotation. If, for instance, semantic clause annotations as we have done them using Coder are to be integrated with syntactic phrase annotations and part-of-speech tagging into one uniform representation, different strata of linguistic organization (syntactic, semantic) and units (clause, phrase, word) have to be merged. Again, this would be feasible simply operating on the different formats, but in a more principled treatment, the units and strata of annotation would have to be defined explicitly in the first place. This can be considered a typical task of document type definition, as handled by, for instance, XML. For this reason we first define a document type definition (DTD) that encodes an annotation grammar suitable for our purposes.

4.1. Multi-layer annotation in XML

The DTD for multi-layer annotation is shown in Figure 9 below.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--DTD generated by XML Spy v3.5 (http://www.xmlspy.com)-->
<!ELEMENT body (seg+)>
<!ELEMENT seg (clause+)>
<!ATTLIST seg
        id ID #REQUIRED
        lang (en | ge) #REQUIRED>
<!ELEMENT clause (phrase | word)+>
<!ATTLIST clause
        semfeat CDATA #IMPLIED>
<!ELEMENT phrase (word+)>
<!ATTLIST phrase
        synform (AdjP | AdvP | NP | PP | VP) #REQUIRED
        synfunc (Adverbial | Complement | Predicate | Subject) #REQUIRED>
<!ELEMENT word (#PCDATA)>
<!ATTLIST word
        pos (ADJA | ADJD | ADV | APPR | ART | AT | CC | CST | DA | DD |
        DD1 | FO | II | JB | JJ | KON | KOUS | MC1 | NE | NN | NN1 |
        NN2 | NP1 | PDAT | PIAT | PIDAT | PPHS2 | REX | RG | RR | VAFIN |
        VBR | VBZ | VM | VV0 | VVPP | YC | YF) #REQUIRED>
```

Figure 9: XML DTD for multi-layer annotation

According to the DTD, text (here called *body*) is split up into the units of clause complex (here: *seg*), *clause*, *phrase* and *word*. The alignment of the source and the target language texts is encoded in the language and ID attributes of each *seg*. Furthermore the unit of the *clause* carries the attribute *semantic feature* (in our case: agentive vs. non-agentive), whereas *syntactic form* and *syntactic function* are encoded as attributes of *phrase*. The attribute *pos* of the unit *word* contains part-of-speech information.
After the format transformations of the different versions of the corpus, the integrated representation of the multiply annotated corpus according to the defined DTD looks as shown in the example in Figure 10.

```
<body>
        <seg id="17" lang="en">
                <clause semfeat="non-agentive">
                        <word pos="CC">and</word>
                    <phrase synform=”NP” synfunc="Subject">
                        <word pos="NN2">textbooks</word>
                    </phrase>
                    <phrase synform=”AdvP” synfunc="Adverbial">
                        <word pos="RR">often</word>
                    </phrase>
                    <phrase synform=”VP” synfunc="Predicate">
                        <word pos="VV0">write</word>
                    </phrase>
                    <phrase synform=”NP” synfunc="Complement">
                        <word pos="DD1">this</word>
                        <word pos="NN1">process</word>
                    </phrase>
                    <phrase synform=”PP” synfunc="Adverbial">
                        <word pos="II">as</word>
                        <word pos="NP1">HCl</word>
                        <word pos="FO">H</word>
                        <word pos="NP1">Cl</word>
                    </phrase>
                        <word pos="YF">.</word>
                </clause>
        </seg>
</body>
```

Figure 10: Integrated XML representation of multiply annotated corpus

The integrated XML-encoded corpus can be validated against the DTD with the help of an XML editor, such as XML Spy[4]. Updates of the DTD can be carried out straightforwardly by automatically generating new DTDs as more data are annotated. The corpus is now in a suitable format to be queried with reference to all its layers of annotation.

4.2. Information extraction

For information extraction a tool is needed which allows querying the corpus with reference to the different layers of corpus annotation. To show how this can work, we employ the MATE system (Mengel 1999, Mengel & Lezius 2000) and use the example of extracting non-agentive NP Subjects (cf. Section 3). Thus, the query needs to refer to the semantic stratum (agentive/non-agentive) as well as the grammatical stratum, and within the latter to a functional category (Subject) and a surface-syntactic category (NP). Figure 11 shows the query formulated in with MATE's query interface.

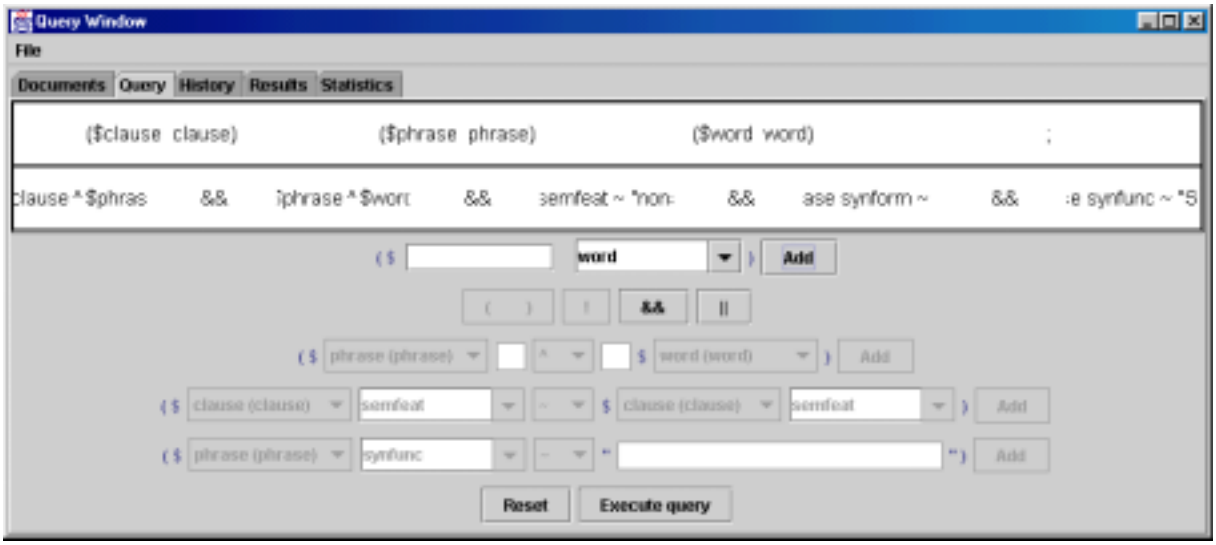

Figure 11: Query for non-agentive NP Subjects with MATE

First of all the elements which are to be included in the query have to be chosen (here: *clause*, *phrase*, *word*). The next step is to define that *clause* governs *phrase* and *phrase* governs *word*. Furthermore the following restrictions are required: The semantic feature of *clause* is *non-agentive*, the syntactic form of *phrase* is *NP* and the syntactic function of *phrase* is *Subject*. The result of this query is shown in Figure 12: two non-agentive NP Subjects ("they", "textbooks") are found in the sample corpus, which occur in the following clauses: "they will readily donate the hydrogen to other molecules" and "textbooks often write this process as HCl H+ + Cl-".

[4] http://www.xml-spy.com

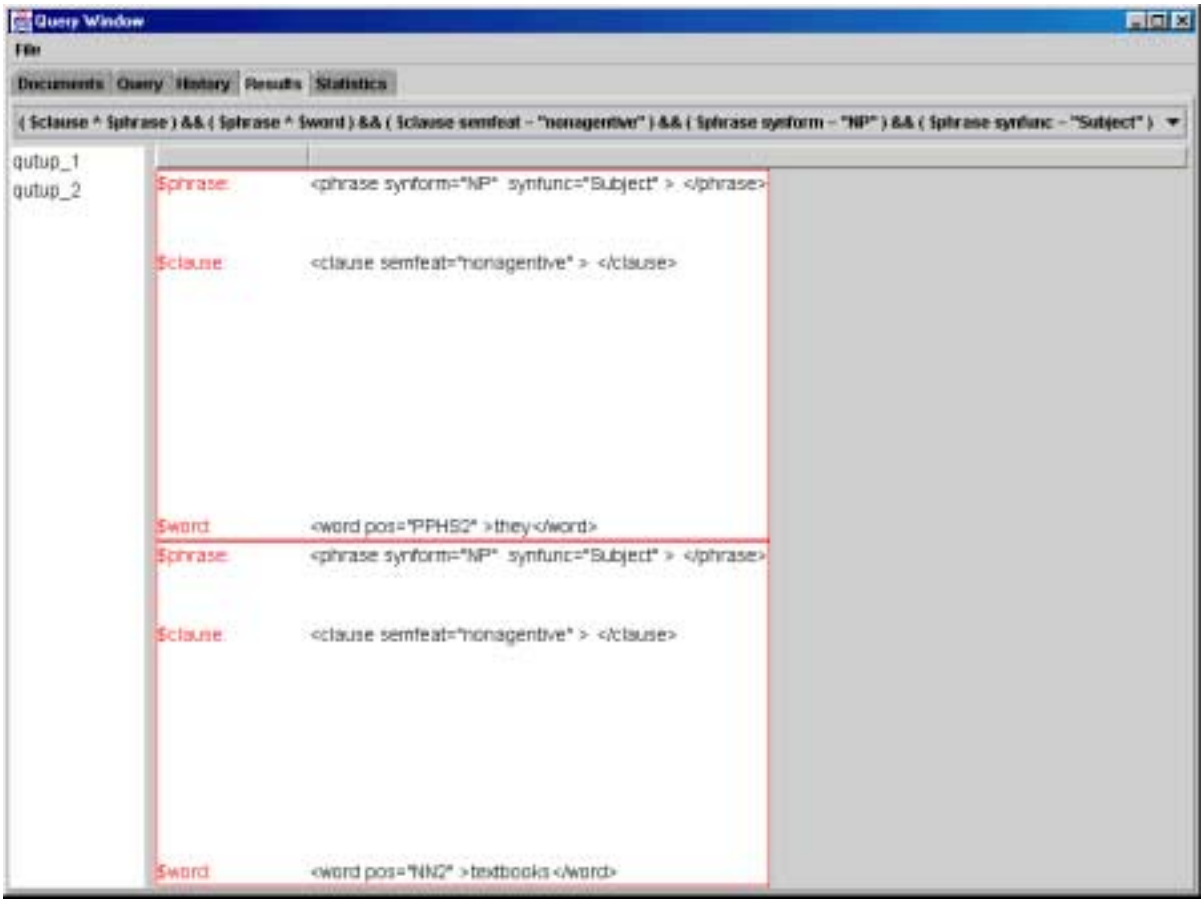

Figure 12: Query results for non-agentive NP Subjects with MATE

5. Summary and conclusions

The present paper has been concerned with the issue of multi-layer linguistic annotation of natural language corpora in the context of the empirical analysis of the specific properties of translations (cf. Sections 1 and 2). the problem that arises here is one of integrating heterogeneous information sources: the techniques used for corpus annotation are all specialized techniques, each catering for one particular kind of annotation (part-of-speech tagging, syntactic parsing, annotation with semantic features) and using one particular representation for the annotated data. In the worst case, a corpus that has been annotated with regard to various kinds of linguistic organization (different kinds of units: words, phrases, clauses; different kinds of strata: syntactic, semantic) will exist in various versions and thus, cross-layer analyses of the corpus will be impeded (cf. Section 3).
As we have shown, it is possible to integrate various layers of annotation, mapping them onto one representation (cf. Section 4). However, at a more abstract level, the problem that is involved here is not merely one of format transformations, but what is needed is a method of integration of heterogeneous information sources. We have suggested that the paradigm of document encoding is one suitable candidate to look at for a solution. Using the Extensible Markup Language (XML), we have presented a grammar of linguistic annotation in the form of an XML-DTD that covers our particular corpus annotation needs, but is adaptable to other kinds of annotation at the same time. Not only can we thus query the corpus making reference to more than one layer of annotation but also, the corpus can be checked for consistency using some standard XML tools such as XML editors, parsers and DTD generators (cf. Section 4). Also, we have shown how information can be extracted from a corpus represented in that way using query languages based on XML. We have illustrated the use of such a query language employing the MATE system. While MATE is a step in exactly the right direction (i.e., an integrated representation of various layers of corpus annotation), at this stage it is only an experimental system and is not fully functional. For instance, for multilingual applications that include translations, parallel concordancing is vital, but does not seem to be fully supported yet.[5]

[5] Note that there is a follow-up project further developing MATE which is about to start, so that a more fully functional system can be expected in due course (http://mate.nis.sdu.dk/).

On a larger scale, it remains to be seen whether new developments in XML-based technology, especially XML-based query languages, such as XML-Query[6] will cater for the needs of the kind of cross-layer corpus analysis we have been concerned with here.

References

- Atril Development SL. 2000. Déjà Vu. Productivity system for translators. Software Manual. (http://www.atril.com/).
- Baker M., 1995. Corpora in translation studies: An overview and some suggestions for future research. In: Target. 7(2): 223-243.
- Baker M., 1996. Corpus-based Translation Studies: The Challenges that Lie Ahead. In: H. Somers (ed.). Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. John Benjamins, Amsterdam & Philadelphia: 175-186.
- Biber D., 1990. Methodological issues regarding corpus-based analyses of linguistic variation. In: Literary and Linguistic Computing. (5): 257-269.
- Brants T., 1999a. Tagging and Parsing with Cascaded Markov Models - Automation of Corpus Annotation. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 6. German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany.
- Brants T., 1999b. Cascaded Markov Models, 1999. In: Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL-99), Bergen, Norway.
- Brants T., 2000a. Inter-Annotator Agreement for a German Newspaper Corpus. In: Second International Conference on Language Resources and Evaluation (LREC-2000). Athens, Greece.
- Brants T., 2000b. TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.
- Brown P.F., J. Cocke, S.A.D. Petra, V.J.D. Pietra, F.J.J.D. Lafferty, E. Mercer & P. Rossin, 1990. A statistical approach to machine translation. In: Computational Linguistics 16(2): 79-85.
- Charniak E., 1993. Statistical Language Learning. The MIT Press.
- Christ O., 1994. A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX 94, 3rd Conference on Computational Lexicography and Text research, Budapest: 23–32 (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/).
- Doherty M., 1993. Parametrisierte Perspektive. In: Zeitschrift für Sprachwissenschaft. 12(1): 3-38.
- Halliday M.A.K., 1985. Spoken and Written Language. Deakin University Press, Geelong, Victoria (Oxford University Press, London, 1989).
- Hansen S. & E. Teich, 2001. Multi-layer analysis of translation corpora: methodological issues and practical implications. To appear in Proceedings of EUROLAN 2001 Workshop on Multi-layer Corpus-based Analysis, Iasi, Romania, July 30 - August 1, 2001.Hansen S., 1999. A contrastive analysis of multilingual corpora (English-German). Diplomarbeit. Universität des Saarlandes, Saarbrücken.
- Hinrichs E., H. Feldweg, M. Boyle-Hinrichs & R. Hauser, 1995. Abschlußbericht ELWIS. Korpusunterstützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik. Technical report, University of Tübingen, Germany.
- Kenny D., 2001. Lexis and Creativity in Translation. A Corpus-based Study. St. Jerome, Manchester.
- Laviosa-Braithwaite S., 1996. The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation. PhD Thesis. UMIST, Manchester.
- Marcus G. P., B. Santorini & M. A. Marcinkiewicz, 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2):313–330, 1993.
- Mengel A. & W. Lezius, 2000. An XML-based representation format for syntactically annotated corpora. In: Proceedings of LREC 2000, Athens: 121-126 (http://mate.mip.ou.dk).
- Mengel A., 1999. Die integrierte Repräsentation linguistischer Daten. In: Gippert J. (ed). Multilinguale Corpora. Codierung, Strukturierung und Analyse (11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung). Prag, enigma corporation: 115-121.
- O'Donnell M., 1995. From Corpus to Codings: Semi-Automating the Acquisition of Linguistic Features. In: Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Stanford University, California: 120-124 (http://cirrus.dai.ed.ac.uk:8000/Coder/index. html).
- Plaehn O. & T. Brants, 2000. Annotate - An Efficient Interactive Annotation Tool. Sixth Conference on Applied Natural Language Processing (ANLP-2000), Seattle, WA.
- Quirk R., S. Greenbaum, G. Leech, J. Svartvik, 1985. A comprehensive grammar of the English language. Longman, London.

[6] http://www.w3.org/XML/Query

- Sager J. C., 1994. Language Engineering and Translation: Consequences of Automation. John Benjamins, Amsterdam.
- Sampson G., 1995. English for the Computer. Oxford University Press, Oxford.
- Scott M., 1996. WordSmith Tools Manual. Oxford University Press, Oxford.
- Sperberg-McQueen C.M. & L. Burnard (eds.), 1999. Guidelines for Electronic Text Encoding and Interchange. Chicago, TEI P3 Text Encoding Initiative. Revised edition.
- Steiner E., 2001. Translations English-German: investigating the relative importance of systemic contrasts and of the text-type „translation“. SPRIK Reports from the project „Languages in Contrast“. University of Oslo, Oslo (http://www.hf.uio.no/german/sprik/english/reports.shtml).
- Teich E. & S. Hansen, 2001. Methods and techniques for a multi-level analysis of multilingual corpora. In: P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (eds.). Proceedings of the Corpus Linguistics 2001 conference. Lancaster: 572-580.
- Teich E., 2001. English - German contrast and commonality in system and text. A methodology for the investigation of parallel and multilingually comparable texts. Habilitationsschrift. Universität des Saarlandes, Saarbrücken.
- Thome G., 1975. Die Übersetzungsprozeduren und ihre Relevanz für die Ermittlung des translatorischen Schwierigkeitsgrads eines Textes. In: W. Wilss. Übersetzungswissenschaft (1). Julius Groos Verlag, Heidelberg.
- Toury G., 1995. Descriptive Translation Studies and beyond. John Benjamins, Amsterdam & Philadelphia.
- W3C-XSLT, 2000. XSL Transformations (XSLT), Version 1.0 (http://www.w3c.org/TR/xslt).
- XCES, 2000. Corpus Encoding Standard for XML. 2000. Vassar College and LORIA/CNRS. (http://www.cs.vassar.edu/XCES/).