Methods and techniques for a multi-level analysis of multilingual corpora

Elke Teich
Institute for Applied Linguistics,
Translation and Interpreting (FR 4.6)
University of Saarland, Germany;
&
Department of Linguistics
University of Sydney, Australia
E.Teich@mx.uni-saarland.de

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

1 Introduction

The present paper discusses the application of a set of computational corpus analysis techniques for the analysis of the linguistic features of translations. The analysis task is complex in a number of respects. First, a multi-level analysis (clause, phrases, words) has to be carried out; second, among the linguistic features selected for analysis are some rather abstract ones, ranging from functional-grammatical features, e.g., Subject, Adverbial of Time, etc, to semantic features, e.g., semantic roles, such as Agent, Goal, Locative, etc.; third, monolingual and contrastive analyses are involved. This places certain requirements on the computational techniques to be employed both regarding corpus annotation and information extraction. We show how a combination of commonly available techniques can fulfil these requirements to a large degree and point out their limitations for application to the research questions raised.

The paper is organized as follows. Section 2 describes the concrete analysis scenario at hand, including the corpus design and the kinds of linguistic features we are interested in extracting from the corpus. Section 3 discusses the application of a range of computational tools in different stages of corpus analysis, ranging from encoding and alignment in the corpus preparation stage over part-of-speech tagging and grammatical and semantic annotation in the linguistic annotation stage to concordance programs in the information extraction stage. Section 4 concludes the paper with a summary.

2 The analysis of a multilingual corpus

One of the primary goals of the kind of corpus analysis described here is to test some hypotheses about the specific features of translations compared to their source language (SL) originals and to comparable original texts in the target language (TL). The language pair we are interested in primarily is English-German. Among the hypotheses tested are Toury's *law of growing standardization* (Toury 1995)/Baker's *normalization* (Baker 1995) and Toury's *law of interference* (Toury 1995). The first says essentially that translations are even more typical of the target language than are original texts in the same language, exaggerating the typical features of the TL (*TL normalization*). The second says that what makes translations a particular kind of text is that the source language always shines through in one way or another (*SL shining-through*).

Given the kinds of relations that are referred to in these hypotheses, we need a corpus that consists of SL originals and their translations as well as comparable original texts in the TL (see Figure 1).

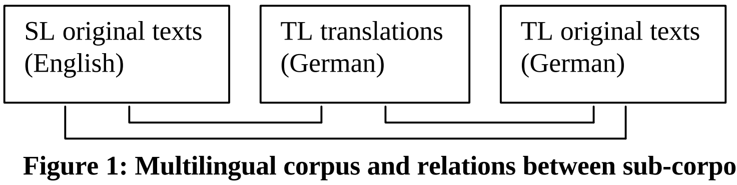


Figure 1: Multilingual corpus and relations between sub-corpora

We make a couple of starting assumptions relating to the kind of analysis that needs to be carried out to test the hypotheses. First, the difference of translations to SL original texts and comparable original TL texts is one of degree, and can thus be measured on a quantitative basis. That is, we can analyze the relations between sub-corpora in terms of frequencies of occurrence of particular linguistic features and compare their distributions across sub-corpora. Second, while a text in a language *l1* and its translation into a language *l2* are comparable simply because they are in a translation relation, for two original texts in a language *l1* and a language *l2* and for two original texts in the same language, we need to make sure that they are comparable using some other criterion. For a definition of the notion of 'comparable', we draw on the concept of register, i.e., linguistic variation according to function in situational context (cf. Quirk et al. 1985; Halliday 1978).

The features selected for analysis are taken from contrastive and monolingual register analysis (Biber 1995; Halliday 1998; Beneš 1981; Fluck 1997). They range from syntactic features such as verb complementation patterns and voice, over semantic features, such as agency, to textual features, such as theme-rheme structure. As will be seen in Section 3, some of these (e.g., passive) can be extracted on the basis of sequences of parts-of-speech, but others (such as agentive) have to be extracted on the basis of manually coded text.

The relation of these features to the testing of the two hypotheses is the following. For example, for the register of scientific writing it is commonly known that English texts from this register are characterized by the frequent use of passives. In German, passive also constitutes a register feature of scientific texts; however, the grammar of German offers other possibilities with similar functions which are "quasi passives" (examples (1) and (2)), so that the core passive occurs less frequently in German original texts than in English original texts of the given register.

(1) *Somit lassen sich auch bei diesen Spielen verschiedene Strategien gegenüberstellen*.
thus let themselves also with these games different strategies oppose
"For these games, too, it would be possible to compare different strategies."
(2) *Dabei ist eine sehr bemerkenswerte Verlagerung der Schwerpunkte zu verzeichnen.*
thereby is a very remarkable shift of emphases to note
"There has also been a remarkable shift in emphasis."

Comparing the frequency of core passives in German translations to their frequency in English originals, two things may happen: Either, there is a significant difference, or there is no significant difference. If there is a significant difference, the frequency of core passives may either be closer to the one in the corpus of SL originals or it may be closer to the one in comparable TL originals. The first would be an indication of SL shining-through, the second an indication of TL normalization.[1]

Syntactic features, such as passive, can be extracted more or less straightforwardly on the basis of text annotated with parts-of-speech. With more abstract features, this is not possible any more. Agency, an attribute of the clause with two values agentive (Agent involved; see example (3)) and non-agentive (no Agent involved; see example (4)) is a case in point. Features such as agency are tested in a similar way as described above for passive, i.e., distributions are compared across corpora and interpreted as SL shining-through or TL normalization.

(3*) She was moving the horses.*
(4) *The horses were moving.*

Finally, if we want to analyze features such as passive or agency according to different registers, we would like to be able to formulate queries such as "Search for all agentive clauses in passive voice in the register of popular-scientific writing". This would require that more than one level of annotation can be referred to at the same time. We will see in Section 3 that this is not trivial with the tools available to date.

Thus, the following requirements are placed on the tools to be used for the kinds of analysis we carry out:

- Syntactic features need to be extracted. Searches on raw text, which can well be successful if one is interested in lexical material, are therefore pointless in the present context. The corpus has to be annotated at least with part-of-speech information so as to enable the extraction of instances of particular syntactic constructions.
- Semantic features need to be extracted. Since semantic analysis cannot be carried out automatically, we need a mechanism for manual annotation, on the one hand, and a query mechanism that is responsive to that annotation.

[1] For details of the methodology and analysis results see (Teich in progress).

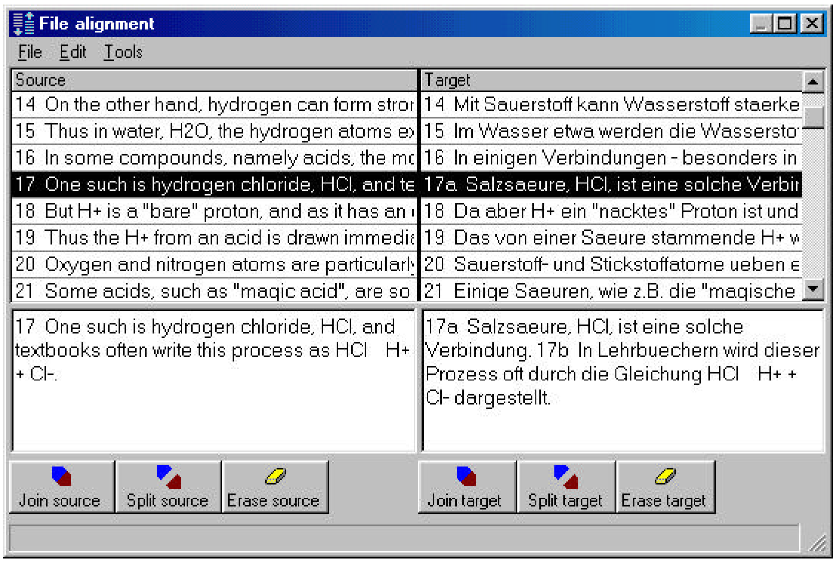
- Contrastive data need to be extracted. Since we need to carry out contrastive analyses, we need the tools to be employed to be applicable for more than one language, on the one hand, and we need querying facilities that are responsive to more than one language at a time.
- Multiply-annotated data needs to be referred to (cf. above).

In the following section we discuss the computational techniques we have employed in corpus preparation, linguistic annotation and for information extraction. The benefits and limitations of each technique for the present analysis requirements are assessed.

3 Computational techniques

3.1 Corpus preparation

Since part of the analysis relates to translations and their SL originals, the parallel corpus needs to be aligned. For this purpose we use the alignment program Déjà Vu.[2] See Figure 2 showing an SL and a TL text aligned with this tool.



Source	Target
14 On the other hand, hydrogen can form stror	14 Mit Sauerstoff kann Wasserstoff staerke
15 Thus in water, H2O, the hydrogen atoms ex	15 Im Wasser etwa werden die Wassersto
16 In some compounds, namely acids, the mc	16 In einigen Verbindungen - besonders in
17 One such is hydrogen chloride, HCl, and te	17a Salzsaeure, HCl, ist eine solche Verbir
18 But H+ is a "bare" proton, and as it has an	18 Da aber H+ ein "nacktes" Proton ist und
19 Thus the H+ from an acid is drawn immedi	19 Das von einer Saeure stammende H+ w
20 Oxygen and nitrogen atoms are particularl	20 Sauerstoff- und Stickstoffatome ueben e
21 Some acids, such as "magic acid", are so	21 Einige Saeuren, wie z.B. die "magische

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 2: Multilingual corpus alignment

Déjà Vu aligns a text and its translation, storing the aligned texts in one file or in two separate files depending on the requirements of the information extraction tool used in later stages of analysis. Files can be exported to translation workbenches and to Microsoft Excel and Access. Figure 3 shows a Déjà Vu output in a TSV (tab separated vector) format.

```
"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."
"18 But H+ is a ""bare"" proton, and as it has an overwhelming attraction to
any electron pair in its vicinity it cannot exist apart from a molecule."
        "18 Da aber H+ ein ""nacktes"" Proton ist und jedes in seiner Naehe
befindliche Elektronenpaar anzieht, kann es nicht wirklich ausserhalb eines
Molekuels existieren."
```

Figure 3: Déjà Vu alignment format

Also, we encode each text of the corpus in terms of a header that provides some meta-information (including title, author, publication, translator, etc) as well as register information (field, tenor, mode).

[2] http://www.atril.com/

Each file is encoded in XML using a modified version of TEI[3] (illustrated in Figure 4) and employing a standard XML editor (here: XML Spy[4]). The text body is annotated for headings, sentences, paragraphs, etc.

```
<tei.2>
    <teiHeader>
        <fileDesc>
            <filename>code_tl_e.txt</filename>
            <subcorpus>popular-scientific (trans_en)</subcorpus>
            <language>English</language>
            <titleStmt>
                <title>Code breaking</title>
                <author>
                    <name>Ewald Osers</name>
                </author>
            </titleStmt>
            <publicationStmt>
                <publisher>The Overlook Press</publisher>
                <pubPlace>Woodstock</pubPlace>
                <date>1999</date>
            </publicationStmt>
            <translation>
                <direction>German-English</direction>
            </translation>
            <sourceText>
                <title>Verschlüsselte Botschaften</title>
                <language>German</language>
                <author>
                    <name>Rudolf Kippenhahn</name>
                </author>
            </sourceText>
            <registerAnalysis>
                <register>popular-scientific</register>
                <field>
                    <experientialDomain>types of code and their function</experientialDomain>
                    <goalOrientation>exposition</goalOrientation>
                    <socialActivity>communication</socialActivity>
                </field>
                <tenor>
                    <agentiveRole>expert to educated layperson</agentiveRole>
                    <socialRole>unequal</socialRole>
                    <socialDistance>maximal</socialDistance>
                </tenor>
                <mode>
                    <languageRole>constitutive</languageRole>
                    <channel>graphic</channel>
                    <medium>written</medium>
                </mode>
            </registerAnalysis>
        </fileDesc>
        <encodingDesc>Modified TEI</encodingDesc>
    </teiHeader>
    <text>
        <body>
         …
        </body>
    </text>
</tei.2>
```

Figure 4: XML corpus encoding

3.2 Linguistic annotation

Depending on how abstract the linguistic features to be analyzed are, the linguistic corpus annotation can be done automatically (using morphological analysis tools, part-of-speech taggers or parsers) or it can be computer-aided (using corpus annotation tools).

A fairly reliable method of syntactic annotation is part-of-speech tagging. The tagger we employ is the TnT tagger, a statistical part-of-speech tagger that analyzes trigrams, incorporating several methods of smoothing and of handling unknown words (Brants 1999). The system is trainable on different languages and comes with the Susanne tagset[5] for English and the Stuttgart-Tübingen tagset[6] for German. It includes a tool for tokenization, which is a preparatory step in the tagging process. In the basic mode, the tagger not only adds a part-of-speech tag to each token, but it omits alternative tags,

[3] http://www.tei-c.org/index.html

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

[5] http://www.cogs.susx.ac.uk/users/geoffs/RSue.html

[6] http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html

together with a probability distribution. It analyzes between 30,000 and 60,000 tokens per second and has an accuracy of about 97 per cent. Figure 5 shows a sample output of TnT, which is in a TSV format.

```
Let            VV0         Coding   VVG
us             PPIO2       keys     NN2
consider       VV0         are      VBR
,              YC          much     DA1
by             II          like     VV0
a              AT1         the      AT
simple         JJ          keys     NN2
illustration   NN1         we       PPIS2
,              YC          use      VV0
the            AT          in       II
problem        NN1         our      APPG
of             IO          daily    JB
depositing     VVG         lives    NN2
a              AT1         .        YF
coding         VVG
with           IW
sender         NN1
and            CC
receiver       NN1
.              YF
```

Figure 5: TnT sample output

When more abstract features are to be coded, annotation has to be carried out manually. Tools supporting such annotation allow the definition of annotation schemes and support manual coding by graphical user interfaces (GUI's). One such tool is Coder (O'Donnell 1995). Coder has five functionalities: chunking-up of texts into units for coding, definition of coding schemes, annotation of texts with a coding scheme, calculating of basic descriptive statistics for a coded corpus and outputting concordances. The chunking mechanism chunks up the text into sentences and paragraphs. If units smaller than sentences need to be annotated, chunking must be done manually. The definition of a coding scheme is supported by a GUI, additions and changes to schemes are straightforward. Coding is supported by another GUI that highlights the unit currently being coded and presents the coding options. The system keeps a record of the codings, on the basis of which a simple descriptive statistics can be calculated and exported. Finally, there is the reviewing function, which is a concordance function operating on the coded features. Figure 6 displays a coding scheme used for coding agency.

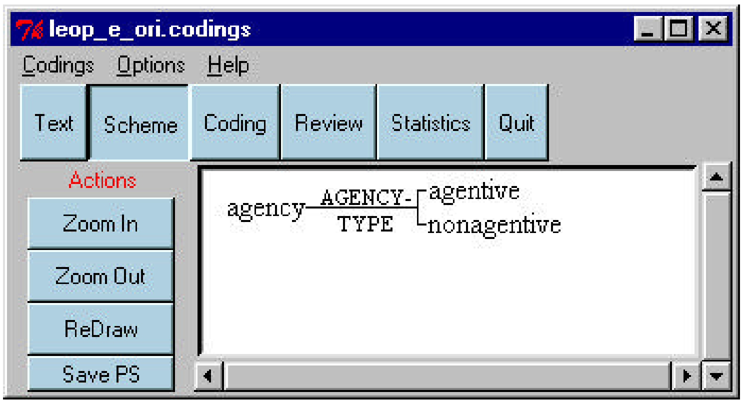


Figure 6: Coder's interface for annotation scheme definition

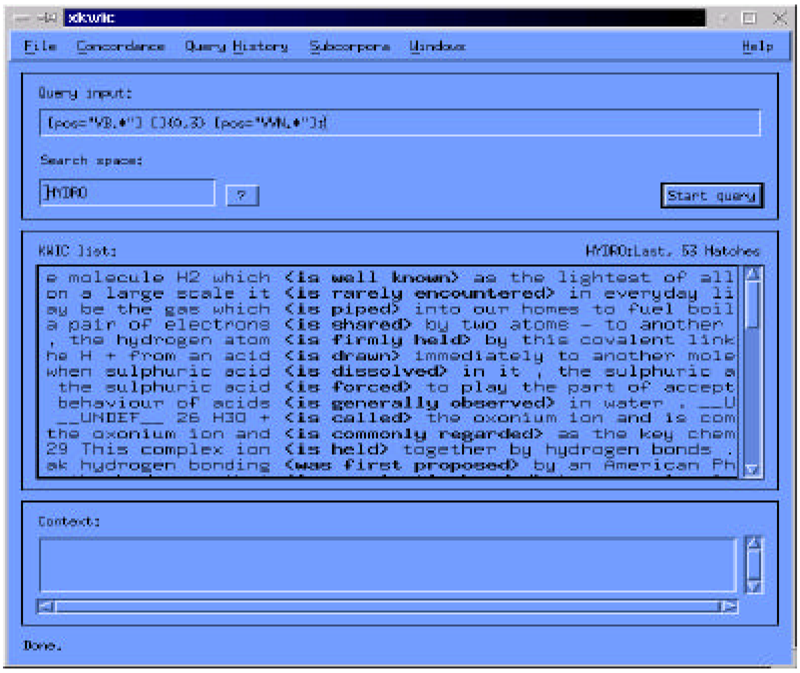
The annotated texts are written out in an XML/SGML-like format. See Figure 7 for an example.

```
<codings>
  <header>
    <scheme root=agency>
         <system name="AGENCY-TYPE" ec="agency" features="agentive nonagentive">
    </scheme>
  </header>
  <body>
   <segment features="agency nonagentive" comment="" ignore=0>The story was about a boy
   </segment>
   <segment features="agency nonagentive" comment="" ignore=0> who had a bucket a net and a dog
   </segment>
   <segment features="agency agentive" comment="" ignore=0> and the dog took the bucket
   </segment>
   <segment features="agency agentive" comment="" ignore=0> and the boy took the net
   </segment>
   <segment features="agency nonagentive" comment="" ignore=0> and they walked over to go to the
                                                             pond to catch a frog
   </segment>
   <segment features="agency nonagentive" comment="" ignore=0>  but when they went
   </segment>
   <segment features="agency nonagentive" comment="" ignore=0> they looked all over almost all day
   </segment>
   <segment features="agency agentive" comment="" ignore=0> but they couldn't find the frog.
   </segment>
  </body>
</codings>
```

Figure 7: Coder output format

3.3 Information extraction

In order to extract particular kinds of linguistic information from the corpus annotated in the ways described above, tools for querying the corpus in terms of the features annotated are needed.
For extracting syntactic information, we use the IMS Corpus Workbench (Christ 1994). The IMS Corpus Workbench is a concordance tool with which it is possible to query for words and/or part-of-speech tags on the basis of regular expressions. Moreover, it allows queries on parallel corpora (aligned translation corpora). The IMS Corpus Workbench consists of two modules: the Corpus Query Processor (CQP) and the user interface (Xkwic). Figure 8 shows Xkwic with a query for passive extraction.



```
[pos="VB.*"] []{0,3} [pos="VVN.*"];
```

HYDRO

HYDRO:Last, 53 Matches

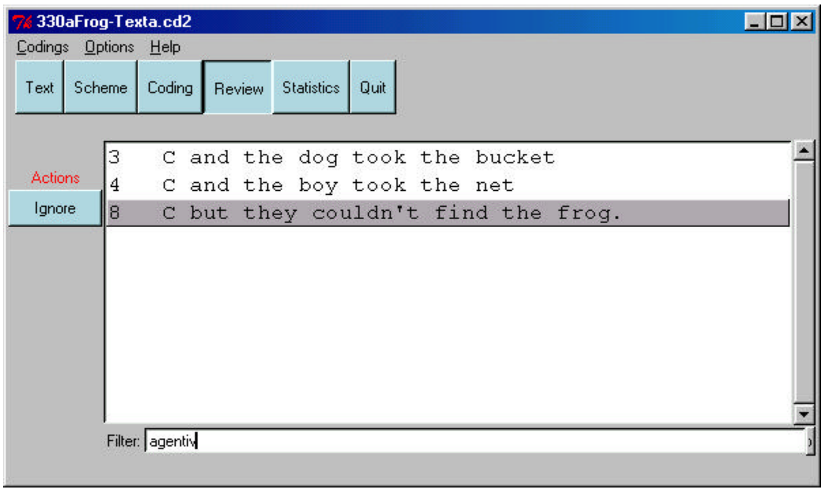
```
e molecule H2 which <is well known> as the lightest of all
on a large scale it <is rarely encountered> in everyday li
ay be the gas which <is piped> into our homes to fuel boil
a pair of electrons <is shared> by two atoms - to another
, the hydrogen atom <is firmly held> by this covalent link
he H + from an acid <is drawn> immediately to another mole
when sulphuric acid <is dissolved> in it , the sulphuric a
 the sulphuric acid <is forced> to play the part of accept
 behaviour of acids <is generally observed> in water . __U
 __UNDEF__ 26 H3O + <is called> the oxonium ion and is com
the oxonium ion and <is commonly regarded> as the key chem
29 This complex ion <is held> together by hydrogen bonds .
ak hydrogen bonding <was first proposed> by an American Ph
```

Done.

Figure 8: User Interface of the IMS Corpus Workbench

The query is based on the part-of-speech tags VB.* (forms of the verb ‘be’) followed by VVN.* (past participle) and zero to three words in between. The results are displayed in the KWIC (keyword in context) list indicating the number of matches as well.

For the extraction of text coded for more abstract features that have been annotated with Coder, the review and statistics functions of Coder can be used for further processing. With the statistics function it is possible to create a descriptive statistics of the analysis; the review function is a concordance function with which it is possible to extract text instances that have been annotated with a particular feature. See Figure 9, which presents an example of extraction of text annotated with the feature agentive.



```
3    C and the dog took the bucket
4    C and the boy took the net
8    C but they couldn't find the frog.
```

Filter: agentiv

Figure 9: Review function of Coder

It is not possible to investigate multilingual corpora with Coder.

4 Summary and conclusions

The analysis task we are faced with in the corpus-based investigation of the linguistic properties of translations places a number of requirements on the computational tools to be used in corpus analysis (cf. Section 2). We have described the application of a set of techniques for the analysis of multilingual corpora ranging from alignment and encoding over linguistic annotation to information extraction (cf. Section 3). Taken together, the tools we have discussed can support the kinds of analysis we carry out, but there are some remaining problems. One has to do with the input and output representations and formats of the individual tools, the other has to do with information extraction.

Each of the tools used for encoding, annotation and extraction employs different input formats that do not necessarily match straightforwardly. While for encoding we have employed XML, 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 from coding are again different across tools: TnT produces a TSV format, Coder produces an XML/SGML-like format. While part of this problem can be dealt with simply by format transformations (e.g., a TSV format can be straightforwardly transformed into an XML format by a Perl script, and an XML-like format can be straightforwardly transformed into XML with the help of XSLT (W3C-XSLT 2000; see Figure 10 for a PoS-tagged text in a simple XML notation), there are some more principled questions involved here to do with the fact that we do multi-level annotation.

```
<head_id_h1> <AT1>A</AT1> <JJ>short</JJ> <NN1>lesson</NN1> <II>on</II> <NN2>keys</NN2>
</head_id_h1>
<p_id_p1> <s_id_s1> <VV0>Let</VV0> <PPIO2>us</PPIO2> <VV0>consider</VV0> <YC>,</YC> <II>by</II>
<AT1>a</AT1> <JJ>simple</JJ> <NN1>illustration</NN1> <YC>,</YC> <AT>the</AT> <NN1>problem</NN1>
<IO>of</IO> <VVG>depositing</VVG> <AT1>a</AT1> <VVG>coding</VVG> <IW>with</IW>
<NN1>sender</NN1> <CC>and</CC> <NN1>receiver</NN1> <YF>.</YF> </s_id_s1> <s_id_s2>
<VVG>Coding</VVG> <NN2>keys</NN2> <VBR>are</VBR> <DA1>much</DA1> <VV0>like</VV0> <AT>the</AT>
<NN2>keys</NN2> <PPIS2>we</PPIS2> <VV0>use</VV0> <II>in</II> <APPG>our</APPG> <JB>daily</JB>
<NN2>lives</NN2> <YF>.</YF> </s_id_s2> <s_id_s3> <VVG>Encrypting</VVG> <VBZ>is</VBZ>
<JJ>similar</JJ> <II>to</II> <VVG>hiding</VVG> AT1>a</AT1> <NN1>message</NN1> <II>in</II>
<AT1>a</AT1> <VVN>locked</VVN> <NN1>box</NN1> <YF>.</YF> </s_id_s3> </p_id_p1>
```

Figure 10: PoS-tagged text in XML format

If, for instance, clause annotations as we have done them using Coder are to be integrated with PoS annotations like the ones given in Figure 10 into one uniform representation, different units of annotation have to be merged. Again, this would be feasible simply operating on the different formats, but in a more principled treatment, the units of annotation would have to be defined explicitly in the first place. This is a typical task of document type definition, as handled by, for instance, XML.

A possible document type definition (DTD) for our annotation purposes could look as displayed in Figure 11.

```
<!ELEMENT Sentence (Clause+)>
<!ELEMENT Clause (Phrase+)>
<!ELEMENT Phrase (Token+)>
<!ELEMENT Token (#PCDATA)>
<!ATTLIST Token
          Pos NMTOKEN #REQUIRED>
```

Figure 11: XML DTD for multi-level annotation

This defines a formal grammar for annotation specifying the units of annotation (sentence, clause, phrase, token) and their attributes (exemplified here with the unit `token` and the attribute `PoS` (part-of-speech). '+' denotes 'one or more occurrences of'.

In information extraction, problems arise for similar reasons. Unless format transformations are carried out (where possible), different tools have to be employed for information extraction and the corpus can only be queried with respect to one level of annotation at a time. It therefore seems to be desirable after all, to have a uniform representation that is built on first principles (see Figure 12 for illustration). Again, this would require employing a document encoding standard in which document types can be properly defined.

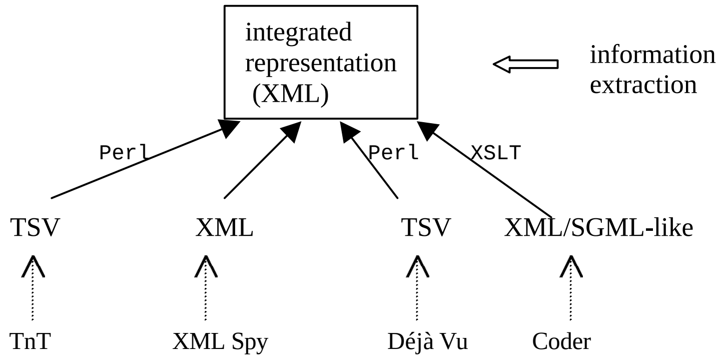


Figure 12: Integrated representation

A recent development in such a direction is the MATE system (Mengel 1999; Mengel and Lezius 2000), which allows for multi-level annotation in a uniform representation, using XML. However, MATE would still need to be tested in a multilingual application of the kind we are involved in here.

Finally, the query mechanisms available in the tools we have tested are either simple Boolean searches on strings (as in the case of Coder) or they are based on regular expressions (as in the case of the IMS workbench). This limits the possibilities of corpus querying. In particular, since the queries in the IMS workbench have to be formulated on sequences of PoS tags, the queries can become quite complex. In our immediate future work, we are going to test more expressive query systems, such as the one implemented in G-Search (Keller et al. 1999), which allows searching with context-free grammars.

To conclude, with a complex corpus analysis task as the one discussed in this paper, we cannot expect to find the *one* ideal tool that can deal with all aspects of annotation and fulfil our particular requirements on information extraction. The concrete tools we have discussed here are exemplars of standard techniques used in corpus linguistics and we would thus expect other linguists with similar analysis requirements to run into the same kinds of problems. Finally, we have formulated the *desideratum* of a uniform representation of corpus annotation, so that on that basis searches on multiple levels of annotation can be carried out. This currently remains an unsolved problem in our analysis scenario.

Acknowledgments

Thanks to Peter Fankhauser and Frank Klein for advice on XML.

References

Baker M 1995 Corpora in translation studies: An overview and some suggestions for future research. *Target* 7(2):223-245.

Beneš E 1981 Die formale Struktur der wissenschaftlichen Fachsprachen aus syntaktischer Hinsicht. In Bungarten T (ed) *Wissenschaftssprache*. München, Fink, pp 185-212.

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

Brants T 1999 *TnT - A Statistical Part-of-Speech Tagger* (User manual). Department of Computational Linguistics, Universität des Saarlandes, Saarbrücken, Germany (http://www.coli.uni-sb.de/~thorsten/tnt/).

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, pp 23–32 (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/).

Fluck H R 1997 *Fachdeutsch in Naturwissenschaft und Technik: Einführung in die Fachsprachen und die Didaktik/Methodik des fachorientierten Fremdsprachenunterrichts*. Heidelberg, Groos.

Halliday MAK 1978 *Language as social semiotic*. Arnold, London.

Halliday MAK 1998 Things and relations: Regrammaticising experience as technical knowledge. In Martin J, Veel R (eds) *Reading Science. Critical and functional perspectives on discourses of science*. London, Routledge.

Quirk R, Greenbaum S, Leech G, Svartvik J 1985 *A comprehensive grammar of the English language*. London, Longman.

Keller F, Corley M, Corley S, Crocker M, Trewin S 1999: Gsearch: A Tool for Syntactic Investigation of Unparsed Corpora. In Uszkoreit H, Brants T, Krenn B (eds) *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, Bergen, pp 56-63.

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, pp 115-121.

Mengel A, Lezius W 2000 An XML-based representation format for syntactically annotated corpora. In *Proceedings of LREC 2000*, Athens, pp. 121-126 (http://mate.mip.ou.dk).

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, pp 120-124 (http://cirrus.dai.ed.ac.uk:8000/Coder/index. html).

Teich E in progress *Contrast and commonality between English and German in system and text. A methodology for investigating the contrastive-linguistic properties of translations and multilingual texts*. Department of Applied Linguistics, Translating and Interpreting, Universität des Saarlandes, Saarbrücken, Germany.

Toury G 1995 *Descriptive Translation Studies and beyond*. Amsterdam, John Benjamins.

W3C-XSLT 2000 *XSL Transformations (XSLT), Version 1.0* (http://www.w3c.org/TR/xslt).