PEDANT: Parallel Texts in Göteborg

Daniel Ridings, *Språkbanken,*
Institutionen för svenska språket, Göteborgs universitet,
S-412 98, Göteborg, Sweden

Abstract: The article presents the status of the PEDANT project with parallel corpora at the Language Bank at Göteborg University. The solutions for access to the corpus data are presented. Access is provided by way of the internet and standard applications and SGML-aware programming tools. The SGML format for encoding translation pairs is outlined together. The methods allow working with everything from plain text to texts densely encoded with linguistic information.

Keywords: SGML, PARALLEL CORPORA, MORPHOSYNTACTIC ENCODING, LEMMATIZATION, MULTIWORD UNITS, COMPOUND WORDS, INTERNET ACCESS

Opsomming: In hierdie artikel word 'n beskrywing gegee van die stand van die PEDANT-projek met parallelle korpora by die Taalbank by die Universiteit van Göteborg. Oplossings vir die verkryging van toegang tot die korpusdata word aangedui. Toegang word verskaf deur middel van die Internet en standaardtoepassings en SGML-sensitiewe programmeringshulpmiddels. Die SGML-formaat vir die enkodering van vertaalpare word gesamentlik geskets. Hierdie metodes laat toe dat gewerk kan word met enigiets vanaf suiwer teks tot tekste wat taalkundig dig geëtiketteer is.

Sleutelwoorde: SGML, PARALLELLE KORPORA, MORFOSINTAKTIESE ENKODERING, LEMMATISERING, MEERWOORDIGE EENHEDE, SAAMGESTELDE WOORDE, INTERNETTOEGANG

Introduction

This is a second progress report dealing with the work being done in Göteborg on parallel texts. It will cast light on new developments and adjust or correct some of the plans expressed in the first report (Danielsson and Ridings 1996b).

Goals

The first of many goals is nearly achieved, namely the creation of a text collection substantial enough to provide raw material for further research. The original ambition to include as many languages as possible has been reduced to concentrating on Swedish, English, German, French and Italian. This language combinations play an important role in the university's translator training programme and in the research interests of graduate students who participate actively in the expansion of the text collection.

A second goal that was deemed to be within realistic reach was to create a foundation for multilingual lexical tools for the translator training program that began at the Faculty of Humanities in January 1997.

These two priorities together with their implications have been the guiding principles to which competing considerations such as target languages, genres, public access, and any attempt at balancing have been subordinated.

Language coverage

So far Swedish has been considered the common denominator for all possible language pairs. This does not necessarily imply that Swedish is always the source language (L1), merely that a text is considered interesting when there is a Swedish version. It is fairly easy to build a collection of texts in German, French and English since those languages are nearly always represented in various documents from the European Union even when the document in question is not available in all of the eleven official languages. Such documents, that is documents without a Swedish counterpart, are not found in the collection.

Despite the fact that Swedish is the pivotal language, this does not exclude the possibility of working with combinations such as French-German, French-Italian, English-French, etc. If a text in Swedish has been aligned with versions in English, French, German, etc., then it is just as possible to align English with French or German for that same text. There is in fact a sizeable amount of material that would permit such investigations. To provide definite numbers to describe the size of a parallel text collection is prone to create misunderstandings. Do the numbers represent the number of words in the whole collection in all the languages or is the collection to be measured by the number of words in only one language for which there are counterparts in another language? In the latter case, is one to count only the words in Swedish texts for which there are always counterparts in one language, or in several languages? Since Swedish is the pivotal language, there will always be a Swedish version of any given text, but not necessarily a German or a French version. After this cautionary note, it can be safely said that there are well over half a million words in Swedish versions for which there exist counterparts in all of the following languages: English, French, German, Italian and Spanish. This means that investigations can be performed on any combination of these languages on a substantial amount of material. If the size is related to language pairs, then there are about a million words of Swedish-English, Swedish-French and Swedish-Italian collections, that is, a million words in Swedish aligned with a million words in English, etc.

The format of PEDANT

The administration of a collection of parallel texts involves all the anticipated

problems and pitfalls associated with monolingual corpora and more. Mistakes were made early on involving the construction of the textual database that need to be rectified. Such mistakes can have far-reaching consequences for the whole system.

PEDANT is represented in two basic formats: a traditional corpus format and a textual database. Both formats have their strengths and weaknesses.

The textual database simplifies the process of providing access to an end user since it is a fairly straightforward process to create a usable graphical interface. The original database manager was Microsoft Access and is still being used. The disadvantage is that this database manager does not lend itself to making material available on the network. This creates a situation where the material is only accessible on one machine and all users must go through the owner of the machine to make queries or printouts. This is often desirable, since not all material is intended to be publicly available, but it is less than satisfactory for the owner of that machine.

The corpus format permits one to use the various tools that have already been developed for monolingual text-processing. One has access to all the standard Unix programs and filters, part-of-speech taggers, lemmatizers, etc.

Both methods of working are manageable as long as only one language pair is involved. The alignment of one unit in one language with another in any given context will always be the same. Once it has been established, it is stable. The parallel text collection in Göteborg, as was mentioned above, contains a substantial amount of material in six languages and the aim is to keep the collection as flexible as possible in order to allow for a diversity of language pairs. Even if the primary concern is with Swedish-French, the possibility of querying German-French for the same text should not be ruled out. This is where the problems arise and where the mistakes were made when the first textual database was created. The following examples can serve to illustrate the problem.

> *** Link: 1-1 ***
>
> I dag genomgår den mänskliga kommunikationen djupgående förändringar genom uppkomsten av globala nätverk för information och kommunikationstjänster, för text, ljud och bilder.
>
> Today, we are experiencing profound changes in human communications with the arrival of global networked information and communication services for text, sound and images.

The point in question here is the alignment. The information content is found in one sentence in Swedish and one sentence in English. Assuming that the information is saved as records or text lines on disk one could easily work out a simple script in any standard Unix language that could search for a word in Swedish: if it is found, then read the next line and the English equivalent will be located there. In such a case each record holds a certain unit of information and is directly related to the following record. So there is a 1-1 relationship between the bits of information and a 1-1 relationship between the orthographic sentences.

When a third language is involved the inappropriateness of this format becomes apparent.

> *** Link: 3-1 ***
> I dag genomgår den mänskliga kommunikationen djupgående förändringar genom uppkomsten av globala nätverk för information och kommunikationstjänster, för text, ljud och bilder. Dessa nätverk innebär spännande utmaningar. De kan också märkbart påverka det språk vi använder.
> Das Auftauchen globaler vernetzter Informations- und Kommunikationsdienste für Text, Ton und Bilder bewirkt heute tiefgreifende Veränderungen in der interpersonellen Kommunikation, die faszinierende Herausforderungen mit sich bringen, gleichzeitig aber auch die Sprachen, die wir benutzen, stark beeinflussen.

In a sense we have the same as the above: one unit of information juxtaposed with the same information in another language. But it now requires three orthographic sentences in the first language and only one sentence in the second. One can still use fairly simple tools but this requires repeating the one version, Swedish in this case, every time that it is aligned with a new language.

If we move our focus from text files on disk to rows and tables in a database the inappropriateness of this format becomes all the more apparent. The various language versions will have to be repeated for every combination with a new language. We will need tables for both the Swedish-English and Swedish-German versions, despite the fact that the Swedish element in both cases is identical. If this is repeated for the other major languages of PEDANT, French, Italian and Spanish, and if to this are added all the combinations of the other languages, French-German, French-Italian, etc., this format becomes quite unwieldy.

This is however easily alleviated by keeping the one version in a separate file from the other, so that instead of two languages in one file, we have two files, each containing one language version. There is still a 1-1 relationship between every line in the two files with regard to content but there is also the possible discrepancy between the orthographic sentences. The easiest way to efficiently access the two pairs is to index the files by word. Each word is given a pointer to the line where it occurs and each line is given a pointer to the equivalent line in the second language version.

This works well for two languages, but when a third one is involved, problems arise. Since the German and English versions quite naturally do not have the same relationship to Swedish with regard to alignment of orthographic sentences they cannot be used against each other. If we want to align them with each other, we will have to create new versions that are correctly aligned with each other, resulting in copy upon copy every time other languages are involved. If we move over from disk files to a database the situation becomes complicated. Even if we were only interested in comparing Swedish to five other languages, the Swedish version will have to be repeated five times.

What is called for is a format that keeps the orthographic sentences stored

in one unit, be it text record or table row, and a separate mechanism that keeps track of which sentence or sentences are aligned with which sentence or sentences in another language. The various text versions can then remain stable with only a small amount of indexing information changing for each language pair. The next two sections will explain how this was done.

Text representation

Independent of the problems mentioned in the previous section, it was obvious that a standard format for the texts would be required, if for no other reason than to provide specifications for software development. At this point "standard" need not refer to anything other than accepted practice within the project.

The PEDANT project decided to use the TEI as the basis for its corpus encoding. The deciding factor was the software package developed by the MULTEXT project consisting of a straightforward API between C programs and SGML encoded files (Thompson et al. 1995). The package is called "Normalised SGML Library" (NSL). This led to a fairly simple solution of the problems described above. PEDANT's first experiences with this package have been described in an earlier report (Danielsson and Ridings 1996a) but NSL has improved since then, partly due to our requests, and our working methods have changed accordingly.

External structure

The lack of information in the literature on parallel texts concerning the issues of representation, access and storage has been pointed out by Armstrong (1996: 17). The present section will be PEDANT's contribution to filling some of these gaps. For this reason it will contain a greater amount of detail than would be expected.

Every language in the PEDANT collection is contained in its own corpus. There is *pedant-se*, *pedant-en*, *pedant-de*, etc.[1] Each corpus is technically a monolingual corpus with no explicit information relating the one language to the other. Each corpus consists of a corpus header followed by the individual texts of that language, each with its own text header. In this respect each corpus is a straightforward implementation of the TEI specifications for corpora. There is no higher element, no project header, encompassing all the corpora.

The SGML elements in the individual texts are very limited: <p> and <s> for the most part, that is, a division into paragraphs and sentences, which is exactly what is required by our aligner. It is technically possible to define other elements as synonyms for <p> and <s> for the aligner, thus allowing for a richer level of annotation, but this is not done to any great extent.[2]

The following two extracts, one from the Swedish corpus and one from the English corpus, are typical.

<P id=se-000001.13><S id=se-000001.13.1 LANG=se>Europeiska rådet noterar med tillfredsställelse några anmärkningsvärda framgångar på området för yttre förbindelser som har uppnåtts sedan dess senaste möte och i vilka Europeiska unionen har spelat en avgörande roll:</S></P>
<P id=se-000001.14><S id=se-000001.14.1 LANG=se>- Undertecknandet av Dayton-avtalet i Paris som sätter punkt för det förödande kriget i det f.d. Jugoslavien och som grundar sig på avsevärda europeiska ansträngningar under de gångna månaderna på det militära och humanitära området samt inom ramen för de förhandlingar som förts; Europeiska rådet erkänner Förenta staternas avgörande bidra vid en ytterst viktig tidpunkt.</S></P>

The equivalent text from the English corpus displays an equivalent structure.

<P id=en-000001.13><S id=en-000001.13.1 LANG=en>The European Council notes with satisfaction some significant achievements in the area of external relations which have occurred since its last meeting and in which the European Union has played a decisive role:</S></P>
<P id=en-000001.14><S id=en-000001.14.1 LANG=en>- the signing in Paris of the Dayton Agreement, which puts an end to the terrible war in former Yugoslavia and builds on considerable European efforts over the preceding months in military, humanitarian and negotiating terms.</S><S id=en-000001.14.2 LANG=en>The European Council recognizes the decisive contribution made by the United States at a crucial moment;</S></P>

Each text, paragraph and sentence is assigned a unique identification number, `id=xx`. The text-id is simply an incrementing number. The sequence is based on Swedish texts. So the first Swedish text is se-000001, the second se-000002, the third se-000003 and so on. It should be recalled that Swedish is the pivotal language so there will be no gaps in the numbering. The corresponding text in the other languages receive the same numbers with prefixes according to the language: de, en, fr, it and es. The numbering for other languages will display gaps, since not every text in Swedish is represented in all the other languages. Leaving gaps in the numbering allows us to integrate such missing texts at a later date. At the same time it is easy to identify which texts correspond to each other since it is seen in the file naming conventions.

The important point to note here is that no alignment information is recorded in these collections. From the technical point of view every language is stored as if it were a monolingual corpus; the SGML *document type* of every collection is `TEICORPUS.2` and can be worked with independently as such. An extract of the corpus header for Swedish can be seen below:

```
<!DOCTYPE teiCorpus.2 SYSTEM "tei2.dtd" [
<!ENTITY % TEI.extensions.ent SYSTEM "pedant.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "pedant.dtd">
]>
<teiCorpus.2>
```

```
<teiHeader  type=corpus>
 <fileDesc>
  <titleStmt>
   <title>PEDANT : Swedish component</title>
   <respStmt>
    <name>Authors</name>
    <resp>Collection and Alignment</resp>
   </respStmt>
   ...
```

The first line declares the document to be a TEI corpus document.[3] The second and third lines contain our customizations of the TEI dtds that we use. We do not use the TEI unaltered but have introduced numerous customizations using the recommended mechanisms for doing so (Sperberg-McQueen and Burnard 1994: 737–744). They are collected in two files that are unique for the project, *pedant.ent* and *pedant.dtd*.

One of the changes we make has to do with the TEI's element <W>. The first thing we do, in the file *pedant.ent*, is to block the TEI's own declaration of this element:

```
<!ENTITY % w 'IGNORE' >
```

There are two things to remember here: (a) Entities get their values on a first come first serve basis and (b) the file *parole.ent* is read and dealt with *before* the relevant dtd in the TEI system. So when the parser arrives at the following declaration in *teiana2.dtd*, the declaration for <W> is ignored:

```
<!ENTITY % w 'INCLUDE' >
<![ %w; [
<!ELEMENT %n.w;   - -   ((#PCDATA | %n.seg; | %n.w; |
                        %n.m; | %n.c;)*)                >
<!ATTLIST %n.w;         %a.global;
                        %a.seg;
          lemma         CDATA                 #IMPLIED
          TEIform       CDATA                 'w'       >
]]>
```

The %w entity has already received a value of "IGNORE" so the whole section is skipped. In the file *pedant.dtd* we have our own project variant of the element as follows:

```
<!ELEMENT %n.w;    - -  ((#PCDATA | %n.seg; | %n.w; |
                        %n.m; | %n.c;)*)                >
<!ATTLIST %n.w;         %a.global;
                        %a.seg;
                        %a.xPointer;
```

```
lemma          CDATA          #IMPLIED
msd            CDATA          #IMPLIED
TEIform        CDATA          'w'          >
```

We added some of our own attributes. We do not use the ANA attribute to provide morphosyntactic tags, but MSD, "morphosyntactic description." There are two reasons for this. In the first place, we do not want to get involved in a complicated mechanism involving IDREFS, which is what ANA expects to be assigned. It is attractive, but complicated, and all the more so since our tags can contain the character "@", which is invalid in those contexts. In the second place, the Parole project (LE2-4017) in which we are involved, uses the MSD attribute, inherited from EAGLES.

In addition to the new attribute for morphosyntactic tags, we also wanted a LEMMA attribute in order to simplify searches and other actions that function best when one has access to lemmata rather than only-word types.

Up to this point our methods of working with the material differ little from those associated with standard monolingual text collections, but the situation changes when our method of providing explicit links between translation equivalents is considered.

Alignment information

For the reasons explained above we decided to keep all details about alignments in separate documents. One document contains all alignments between two languages. There is a document for Swedish-English and another document for Swedish-German and if we ever decide to align English with German we would have yet another document for English-German. Each one of these documents contain all of the alignment information for all the texts for the language pair it describes.

The way we do this is by creating a new "corpus", technically speaking, but the individual "texts" in the corpus document (the SGML corpus document) do not contain any natural language, only alignment links. For any given document from the monolingual collection that has been aligned with another language there will be the same number of "paragraphs" and the same number of "sentences", that is, the same number of <P> and <S> elements, but their content will be minimal.

We have introduced two new elements to the TEI system, <SSEG> and <TSEG>, "source segments" and "target segments." We see the results of alignment in pairs; one segment of one text is aligned with one segment of another text. These segments can contain a combination of "sentences". The source segment might consist of two sentences while the target segment consists of only one. The possible combinations are: 1-1, 2-1, 2-2, 1-0 and vice versa. An alignment "pair", consisting of exactly one <SSEG> and one <TSEG>, is itself contained in one single <SEG>, which should possibly be renamed to <ASEG>, alignment segment, by analogy.

So a paragraph can contain one or more <SEG> elements, which each contain an alignment pair, <SSEG> and <TSEG>, and each of these latter elements can contain one or more sentences which in turn can contain one or more words and so on down the branches of the SGML document tree, i.e.:

```
<P>
   <SEG>
        <SSEG></SSEG>
        <TSEG></TSEG>
   </SEG>
</P>
```

Information is attached to the <SSEG> and <TSEG> elements that points back into the relevant monolingual corpora where the actual sentences are found. For example, an alignment in the Swedish-English collection appears as follows:

```
<P ID=SE-000001.13>
<SEG ID=SE-000001.S13>
<SSEG DOC=sedoc FROM='id (SE-000001.13.1)' ID=SE-000001.SS13></SSEG>
<TSEG DOC=endoc FROM='id (EN-000001.13.1)' ID=EN-000001.TS13></TSEG>
</SEG>
</P>
<P ID=SE-000001.14>
<SEG ID=SE-000001.S14>
<SSEG DOC=sedoc FROM='id (SE-000001.14.1)' ID=SE-000001.SS14></SSEG>
<TSEG DOC=endoc FROM='id (EN-000001.14.1)'
                TO='id (EN-000001.14.2)' ID=EN-000001.TS14></TSEG>
</SEG>
</P>
```

The various elements for "segments" receive their own ID values, since they do not occur anywhere else than in this document. The <SSEG> and <TSEG> elements have additional attributes, DOC, FROM and TO. Alignment information is assigned to these attributes.

The DOC attribute provides information on the corpus in which the sentences can be found, that is, it points to the relevant monolingual corpus.

The FROM and TO attributes contain the extent, measured in sentences, in the monolingual corpus. The values they are assigned, are the values that have been assigned to the ID attribute of the sentences in question. A missing TO attribute defaults to the same as the FROM attribute, that is, one sentence. In the first paragraph in the example above, we have a 1-1 alignment. In the second paragraph we have a 1-2, that is, one Swedish sentence corresponds to two English sentences.

Similar information is recorded for every aligned text in PEDANT. The information is stored as its own corpus, that is, a Swedish-English corpus, a Swedish-German corpus, etc. The corpus header is as follows:

```
<!DOCTYPE teiCorpus.2 SYSTEM "tei2.dtd" [
<!ENTITY % TEI.extensions.ent SYSTEM "pedant.ent">
<!ENTITY % TEI.extensions.dtd SYSTEM "pedant.dtd">
<!ENTITY sedoc SYSTEM "/Pedant/Corpus/Swedish/Swedish.nsg" CDATA SGML>
<!ENTITY endoc SYSTEM "/Pedant/Corpus/English/English.nsg" CDATA SGML>
]>
<?NSL LINKS S SSEG S TSEG>
<teiCorpus.2>
 <teiHeader  type=corpus>
  <fileDesc>
   <titleStmt>
    <title>PEDANT : Swedish-English component </title>
    <respStmt>
     <name>Authors </name>
     <resp>Collection and Alignment</resp>
    </respStmt>
    ...
```

The beginning of the header contains three significant lines: 4 and 5, and 7. Recall the value assigned to the DOC attributes of the <SSEG> and <TSEG> elements: sedoc and endoc. Lines 4 and 5 map these to the relevant monolingual corpora. Line 7 is a *processing instruction* and is unique for the LT NSL package we are using as mentioned above. It provides the information that <S> elements can be linked to <SSEG> and <TSEG> elements.

All of this is tied together via the LT NSL utility called *mknsg*. We run the command as follows:

```
mknsg -D sgml -D sgml/pedant Swedish-English.sgm | <further processing>
```

This command ties all the information together and prints to standard output where it can be piped into other programs for further processing or redirected to disk. The output is as follows:

```
<P ID=SE-000001.13>
<SEG ID=SE-000001.S13>
<SSEG ID=SE-000001.SS13>
<S ID=SE-000001.13.1 LANG=SE>Europeiska rådet noterar med tillfredsställelse
några anmärkningsvärda framgångar på området för yttre förbindelser som har uppnåtts
sedan dess senaste möte och I vilka Europeiska unionen har spelat en avgörande
roll:</S></SSEG>
<TSEG ID=EN-000001.TS13>
<S ID=EN-000001.13.1 LANG=EN>The European Council notes with satisfaction
```

```
some significant achievements in the area of external relations which have occurred since
its last meeting and in which the European Union has played a decisive role:</S>
</TSEG>
</SEG>
</P>
<P ID=SE-000001.14>
<SEG ID=SE-000001.S14>
<SSEG ID=SE-000001.SS14>
<S ID=SE-000001.14.1 LANG=SE>- Undertecknandet av Dayton-avtalet i Paris
som sätter punkt för det förödande kriget i det f.d. Jugoslavien och som grundar sig på
avsevärda europeiska ansträngningar under de gångna månaderna på det militära och hu-
manitära området samt inom ramen för de förhandlingar som förts; Europeiska rådet er-
känner Förenta staternas avgörande bidrag vid en ytterst viktig tidpunkt.</S>
</SSEG>
<TSEG ID=EN-000001.TS14>
<S ID=EN-000001.14.1 LANG=EN>- the signing in Paris of the Dayton Agree-
ment, which puts an end to the terrible war in former Yugoslavia and builds on consid-
erable European efforts over the preceding months in military, humanitarian and nego-
tiating terms.</S>
<S ID=EN-000001.14.2 LANG=EN>The European Council recognizes the deci-
sive contribution made by the United States at a crucial moment;</S></TSEG>
</SEG></P>
```

This might seem to be a complicated procedure at first sight, but most of the details are fully automated and there are some further advantages in that all software can build upon the same library. This will be illustrated by our lemmatizer, PEDAL.[4]

The expansion of the links, as mentioned above, can be saved to disk or piped into other SGML tools. The LT NSL package has introduced a way of working with "semivalid sgml". A valid SGML file is run through the basic tool *mknsg*, which parses the DTD permanently and caches a binary version on disk. This cached version is the one used by all subsequent programs. One of the characteristics of "semivalid sgml" is that the segment piped into a program must be valid when compared to the cached DTD, but the segment piped in need not itself contain the whole document tree. The excerpt above, for instance, would be valid input to other NSL tools, since the elements and their contents are valid. There is no header, no <TEXT> elements, but the <P> elements are syntactically correct with regard to that part of the DTD that deals with them.

One of the tools, *sggrep*, that comes along with the library can help to exemplify this. Let us assume that we want to identify all translation pairs in which the Swedish half contains the expression "sätter punkt för". To do so, we would replace the <further processing> in the above command with:

```
sggrep ".*/SEG" "SEG/SSEG" "sätter punkt för"
```

The first parameter provides the program with the depth of our query in the SGML document tree, namely, down to the level of <SEG>. The second parameter provides the subquery, the space in which the search will be done. In effect, it defines the segment of the SGML document we want reported from the query. Recall that alignment pairs are contained in the <SEG> element, exactly one pair per element. So the result returned will be the whole <SEG> element. The actual query, however, will be limited to the <SSEG> element. The program will not look for "sätter punkt för" in the English element, <TSEG>, but it will be returned together with the Swedish segment, since it too is contained by the <SEG> element. The result is similar to what appears in the excerpt above and is, in fact, the way it was produced for this report. It could have been piped into yet other tools performing statistical analysis, lemmatization, etc. The possibilities are numerous when all tools work with the same API. Further examples can be found in an earlier report (Danielsson and Ridings 1996a: 7–12).

Database

The original database manager was Microsoft Access and it is still being used for quite a few tasks successfully. Since we wanted to provide network access as well, we began experimenting with some of the freely available relational databases for Unix: miniSQL and mySQL.

Network access

Network access is provided by way of (a) miniSQL and more recently mySQL, (b) cgi-scripts and (c) any browser on any platform that supports tables and forms. Those interested can turn to *http://svenska.gu.se/PEDANT* for a demonstration.

Some of our first attempts at identifying equivalents on the basis of one Swedish word can be seen in figure 1 in the Appendix. It is a web-based system working against the mySQL database.

Linguistic tools

The tagset

PEDANT uses a tagset that is mappable on a 1-1 basis with the SUC tagset. The SUC tagset was designed by Eva Ejerhed. In February 1997 Ejerhed and Ridings adjusted the PAROLE tagset in such a way that the PAROLE tagset and the SUC tagset are interchangeable. This is evidenced in the SGML version of the SUC corpus. A table comparing the two sets together with example words can be found at *http://ldb20.svenska.gu.se*.

Unlike our lemmatizer, Brill's tagger has not yet been made SGML aware. The tagger and the alignment program are the only two programs left in our repertoire that have not been written for the NSL API. In the case of Brill's tag-

ger we feel that there are other adjustments we want to make, particularly in the lexical rule component, but the time or human resources are not yet available.

This is not so problematic since Brill's tagger requires two properties of a text that is to be tagged: (a) it must be segmented into sentences and (b) it must be tokenized.

Our texts must be segmented into sentences before they can be aligned so the first requirement is met. With regard to the second requirement, we have a tokenizer that works directly with the SGML files through the NSL API. This means that we can export our texts from SGML in such a way that there is a 1-1 relationship between the exported file's sentences and tokens and the original texts. This being the case it is a simple matter to take the tagged results from Brill's tagger and map the morphosyntactic tags back onto the proper attributes of each token.

The results of tagging are as follows:

```
<P ID=SE-000001.13>
<S ID=SE-000001.13.1 LANG=SE>
<W MSD='AQP0SNDS'>Europeiska</W>
<W MSD='NCNSN@DS'>rådet</W>
<W MSD='V@IPAS'>noterar</W>
<W MSD='SPS'>med</W>
<W MSD='NCUSN@IS'>tillfredsställelse</W>
<W MSD='DI@0P@S'>några</W>
<W MSD='AQP0SNDS'>anmärkningsvärda</W>
<W MSD='NCUPN@IS'>framgångar</W>
<W MSD='SPS'>på</W>
<W MSD='NCNSN@DS'>området</W>
<W MSD='SPS'>för</W>
<W MSD='AQC00N0S'>yttre</W>
<W MSD='NCUPN@IS'>förbindelser</W>
<W MSD='PH@000@S'>som</W>
<W MSD='V@IPAS'>har</W>
<W MSD='V@IUPS'>uppnåtts</W>
<W MSD='RG0S'>sedan</W>
```

A file that is morphosyntactically tagged is then run through PEDAL, the lemmatizer, assigning one of three alternatives to the LEMMA attribute of the <W> element: (a) a lemma if a lemma is found with a matching morphosyntactic tag, (b) "not-found" if it was not possible to resolve the word type to its base form, or (c) "no-msd-match" in the event that the word type could be resolved to a base form but the morphosyntactic description associated with the word type does not match the description provided by PEDAL.

PEDAL works directly with SGML files and provides an opportunity to illustrate how simple it is to integrate an SGML parser with one's own code. The core of the lemmatizer is made up of the following lines of code:

```
strcpy(qustr, ".*/W");
qu=ParseQuery(dct, qustr);

while( ( item=GetNextQueryItem(inf, qu, outf ) ) ) {
    msdVal = GetAttrStringVal(item, AttrName);
    strcpy(wordtype, item->data->first);
    if (msdVal != NULL) {
        if (*msdVal == 'N'
            || *msdVal == 'V'
            || *msdVal == 'A' || *msdVal == 'D'
            || *msdVal == 'P') {
            strcpy(lemma,lemmatize(wordtype, msdVal));
        } else {
            strcpy(lemma, item->data->first);
        }
        if (!strcmp("not-found", lemma)) {
            strcpy(lemma,guess_lemmatize(wordtype, msdVal));
        }
        PutAttrVal(item, LemmaAttr, lemma);
    }
    PrintItem(outf, item);
    FreeItem(item);
}
```

The first two lines set up the SGML query, `".*/W"`. The dot is a wildcard meaning "any" and the star is the standard kleene star meaning "zero-or-more". What is being referred to here are elements in the SGML document tree. On the fourth line the query can be read as "search down the SGML tree, traversing all elements until we descend down to the level of <W>". That is the base element of our documents. There are no other elements below the <W> element. The API passes all elements down to that level to the output, but turns over <W> elements to the program for processing. After processing this base element is printed to output by the third line from the end, `PrintItem(outf, item)`, thus completing the whole document.

The fifth line:

```
msdVal = GetAttrStringVal(item, AttrName);
```

reads the MSD attribute, the morphosyntactic tag, of each <W> element. This is passed on to the lemmatizing routine. At this point we only lemmatize nouns, verbs, adjectives, determiners and pronouns. All other classes of words simply get their word type copied to the LEMMA attribute (line 14).

The word type and the morphosyntactic tag are then sent to the lemmatizing routine (line 12). The core of the lemmatizing routine is:

```
resp = recognizer(wordtype,  Lang, 0, 0, (FILE *)NULL);
strcpy(lemma, "not-found");
if (resp != (RESULT *)NULL) {
    strcpy(lemma, "no-msd-match");
    for (rp=resp; rp; rp=rp->link) {
        sscanf((char *)rp->feat, "\[%s %s", tmp_l, tmp_a);
        eos = (char *)rindex(tmp_a, '\0');
        --eos; *eos = '\0';
        if (!strcmp(tmp_a, msdtag)) {
            strcpy(lemma, tmp_l);
        }
    }
    free_result(resp);
}
return(lemma);
```

The first line searches for all the possible base forms of the word type. Before anything else is done, the lemma is set to "not-found". If it is found, then this will be overwritten, if it is not found, then this will be returned. In line 3 a check is made for results: if there were results, then the lemma is set to "no-msd-match", that is, a lemma was identified, but its morphosyntactic description did not match the word type's. This will be overwritten if it proves not to be the case.

The `for`-loop walks through all of the possible interpretations of the word type. The `if`-statement in the loop compares each interpretation's morphosyntactic description with that of the original word type. If they match, then the proposed lemma is copied to the lemma string and will eventually be returned. If no matches are found, then the lemma string retains "no-msd-match" and this is returned. This signals all the places where the results of the tagger deserve manual control. There can be other places as well, but this catches a lot of the mistakes, though in practice, they are not that many. The resulting output is as follows:

```
<P ID=SE-000001.13>
<S ID=SE-000001.13.1 LANG=SE>
<W LEMMA='europeisk' MSD='AQP0SNDS'>Europeiska</W>
<W LEMMA='råd' MSD='NCNSN@DS'>rådet</W>
<W LEMMA='notera' MSD='V@IPAS'>noterar</W>
<W LEMMA='med' MSD='SPS'>med</W>
<W LEMMA='tillfredsställelse' MSD='NCUSN@IS'>tillfredsställelse</W>
<W LEMMA='någon' MSD='DI@0P@S'>några</W>
<W LEMMA='anmärkningsvärd' MSD='AQP0SNDS'>anmärkningsvärda</W>
<W LEMMA='framgång' MSD='NCUPN@IS'>framgångar</W>
<W LEMMA='på' MSD='SPS'>på</W>
<W LEMMA='område' MSD='NCNSN@DS'>området</W>
<W LEMMA='för' MSD='SPS'>för</W>
```

```
<W LEMMA='yttre' MSD='AQC00N0S'>yttre</W>
<W LEMMA='förbindelse' MSD='NCUPN@IS'>förbindelser</W>
<W LEMMA='som' MSD='PH@000@S'>som</W>
<W LEMMA='ha' MSD='V@IPAS'>har</W>
<W LEMMA='uppnå' MSD='V@IUPS'>uppnåtts</W>
```

Three lines in the code have not been discussed yet:

```
if (!strcmp("not-found", lemma)) {
    strcpy(lemma,guess_lemmatize(wordtype, msdVal));
}
```

As mentioned above, if the lemmatizing routine does not find a match, it returns the string "not-found". It turns out that almost all of these cases are compound words. This is a familiar problem for all of those dealing with Germanic languages (Hellberg 1978: 21–28; Karlsson 1992: 15–17) and renders otherwise attractive publicly available packages more or less useless.

Our approach to this is simple but works satisfactorily. We have a limited lexicon section that allows certain forms to lead back into the main lexica (cf. Karlsson 1992: 15). In general, however, we are working with the assumption that (a) "not-found" words are compounds and (b) that the longest segment on the right-hand side of the compound for which a base form is identified with the correct morphosyntactic description, provides us with the best compound boundary. In other words, the routine `guess_lemmatize` walks through the word type backwards and returns the longest segment.

The words that have been identified by the second form of the lemmatizing routine can be easily identified, i.e.:

```
pedal Swedish.msd.nsg | sggrep -r ".*/W" "W[LEMMA='.*_.*']" ""
```

The above command pipes the result of the lemmatizer into one of the tools that is included in the NSL package, *sggrep*, a version of *grep* that understands the structure of SGML documents.

The first parameter of the command, `".*/W"`, provides the depth of the query, that is, all the way down the document tree to the level of the `<W>` element. The second parameter, `"W[LEMMA='.*_.*']"`, provides the subquery, that is, the scope of the document that will be queried and returned. A search on attributes to an element is provided within square brackets, the `LEMMA`, in this case. The third parameter is empty because we are not searching for specific words (element content), but for words with certain attributes. Attribute values can be expressed with regular expressions if the `-r` flag is provided. So we are searching for all lemmata which contain an underscore, put there to mark the suggested compound boundary. The result is:

```
<W LEMMA='regerings_konferens' MSD='NCUSN@DS'>regeringskonferensen</W>
<W LEMMA='reflexions_grupp' MSD='NCUSG@DS'>reflexionsgruppens</W>
```

```
<W LEMMA='morgon_dag' MSD='NCUSG@DS'>morgondagens</W>
<W LEMMA='Dayton-_avtal' MSD='NCNSN@DS'>Dayton-avtalet</W>
<W LEMMA='åtgärds_plan' MSD='NCUSN@DS'>åtgärdsplanen</W>
<W LEMMA='Barcelona_förklaring' MSD='NCUSN@DS'>Barcelonaförklaringen</W>
<W LEMMA='medelhavs_område' MSD='NCNSN@DS'>medelhavsområdet</W>
<W LEMMA='medelhavs_område' MSD='NCNSN@DS'>Medelhavsområdet</W>
<W LEMMA='åsikts_utbyte' MSD='NCNSN@IS'>åsiktsutbyte</W>
<W LEMMA='Europa_parlament' MSD='NCNSG@DS'>Europaparlamentets</W>
<W LEMMA='diskussions_fråga' MSD='NCUPN@DS'>diskussionsfrågorna</W>
<W LEMMA='valuta_enhet' MSD='NCUSN@DS'>valutaenheten</W>
<W LEMMA='anpassnings_kostnad' MSD='NCUPN@DS'>anpassningskostnaderna</W>
<W LEMMA='råds_förordning' MSD='NCUSN@IS'>rådsförordning</W>
<W LEMMA='ecu-_korg' MSD='NCUSN@DS'>ecu-korgen</W>
<W LEMMA='valuta_enhet' MSD='NCUPN@IS'>valutaenheter</W>
<W LEMMA='Ekofin-_råd' MSD='NCNSN@DS'>Ekofin-rådet</W>
<W LEMMA='euro-_sedel' MSD='NCUPN@DS'>euro-sedlarna</W>
<W LEMMA='Budget_disciplin' MSD='NCUSN@DS'>Budgetdisciplinen</W>
<W LEMMA='budget_ordning' MSD='NCUSN@DS'>budgetordningen</W>
<W LEMMA='euro-_område' MSD='NCNSN@DS'>euro-området</W>
```

One of the above analyses is the result of a previous correction of PEDAL's lexicon, namely:

```
<W LEMMA='medelhavs_område' MSD='NCNSN@DS'>medelhavsområdet</W>
```

The original analysis put the compound boundary after "medel" (middle), since "havsområde" (sea-area) is listed in the lexicon making it the longest right-hand segment. We added "medelhavs" to the lexical listings with a continuation into the noun lexicon. This results in the compound being recognized and returned instead of "not-found" and the guessing routine never gets called. All other forms in the above list have been produced by the principle of longest segment to the right with matching MSD values of the original word type.

Excursus

The method of storing alignment information was outlined above. At this point one of the benefits of our architecture can be indicated by drawing attention to the following lines:

```
<!ENTITY sedoc SYSTEM "/Pedant/Corpus/Swedish/Swedish.nsg" CDATA SGML>
<!ENTITY endoc SYSTEM "/Pedant/Corpus/English/English.nsg" CDATA SGML>
```

These lines linked back into the individual monolingual corpora for the language pair that was aligned. The versions of *Swedish.nsg* and *English.nsg* in these lines are the versions with a minimum of mark-up: paragraphs and sentences.

Let us assume that the lemmatization described in the previous section was saved to disk, rather than just piped into other tools, for example to *Swedish.lemma.msd.nsg*. We can then change the two lines above to read:

```
<!ENTITY sedoc SYSTEM
            "/Pedant/Corpus/Swedish/Swedish.lemma.msd.nsg" CDATA SGML>
<!ENTITY endoc SYSTEM "/Pedant/Corpus/English/English.nsg" CDATA SGML>
```

This provides us with the same information about alignment pairs, but this time we will have access to a richer array of information when it comes to the Swedish component, namely lemmata and morphosyntactic descriptions. The paragraph and sentence IDs are the same in both versions, the base version and the morphosyntactic tagged version, so the alignment information in the Swedish-English component will point back to the same sentences.

This method of working allows us to experiment with various levels of annotation without cluttering up the base version. One level of annotation that interests us at the moment is one that marks phrases below the sentence level. This will be dealt with below.

Current directions

This section is only a preliminary sketch of some of the directions that are being pursued at the moment. It is by no means exhaustive since there are now many others involved with various investigations and they will be reported on in their own time by the individual researchers involved.

Equivalents below the sentence level

Once a collection of orthographical sentences has been correctly aligned with translation equivalents the most natural next step is to identify smaller chunks. Word-to-word alignment will always be difficult since most translations do not display such a structure.[5]

The department's background is lexicographical and the work being done in PEDANT reflects that fact. One of our distant goals is to create a lexicographical workbench for bilingual lexicography. A major aspect of this goal is to introduce the methods from corpus-based monolingual lexicography to the multilingual sphere.

Word tuples

N-grams models of word-tuples are basically lists of word-pairs, word-triples, word-quadruplets and so on that are provided with frequency and likelihood information (Atwell 1996: 160). Likelihood, in this section, is based on the likelihood ratio in Dunning (1994).[6]

Lists are created of all bigrams, trigrams and quadrigrams in a subset of the Swedish-English component of PEDANT.[7] For examples of Swedish and English trigrams see figures 2 and 3 respectively in the Appendix.

The motivation for performing the tests was to see if "phrases", in a loose sense, in one language showed any inclination to be translated by phrases in another language. There cannot be that many translations of "as soon as possible" into Swedish other than "så snart som möjligt" and there might even be other hidden combinations of words that do not directly come to mind.

The purpose, at this stage, is not to identify which phrases are translations of each other. The tables are sorted in descending frequency according to the significance of the "tuple" in its own language, and the equivalent in the other language, when compared to its own corpus of words, cannot be expected to show the same degree of equivalence. There is simply no connection between the two with regard to the ranking the phrases get in their respective languages.

The assumption is that a phrase in one language might be translated by a phrase in another. If this is the case, then, should we search in the parallel texts, have a translation pair in front of us and find a phrase in L1 and also in L2, it is worth noting if the phrase in L2 is a translation of the phrase in L1. This method might succeed if the language pairs are not overloaded with phrases, which seems to be a fair assumption to begin with, considering the fact that the vast majority of alignments are 1-1, that is, one sentence to one sentence.

The "phrases" identified by the $-2 \log \lambda$ formula have been automatically tagged with a global search-and-replace command so that they are enclosed in the <PHR>...</PHR> chunk. This was done by taking the first 75 trigrams and thereafter the first 75 qaudrigrams.[8]

As mentioned above, this is just experimental and not a full report, but the first indications are encouraging. In the excerpt below we see the results of searching for the English phrase "with regard to." The following is the result of piping our corpus of pure links as described above through the *sggrep* utility:

```
mknsg -D sgml -D sgml/pedant Swedish-English.sgm \
  | sggrep ".*/SEG" "SEG/.*/TSEG/.*/PHR" "with regard to"
```

— in other words, search for the phrase "with regard to" only in the English half of the translation pairs.

> <SEG ID=SE-000001.S119>
> <SSEG ID=SE-000001.SS119><S ID=SE-000001.91.1 LANG=SE>- förbättra deras finansiella miljö genom en förbättrad tillgång till kapitalmarknaderna och främja utvecklingen av europeiska investeringsfondens funktion <PHR>**när det gäller**</PHR> <PHR> små och medelstora</PHR> företag.</S></SSEG>
> <TSEG ID=EN-000001.TS119><S ID=EN-000001.91.1 LANG=EN>- improve the financial environment for them by means of better access to capital markets and encourage development <PHR>**of the European**</PHR> Investment Fund function <PHR> **with regard to**</PHR> SMEs.</S></TSEG>

</SEG>
<SEG ID=SE-000004.S345>
<SSEG ID=SE-000004.SS345><S ID=SE-000004.196.2><PHR>**Små och medelstora företag**</PHR>, som utgör nästan hälften av den ekonomiska basen, möter speciella svårigheter och särskilt i <PHR>**frågor som rör**</PHR> finansiering (t.ex. den effektiva räntan är ofta 2 till 3 punkter högre än i utvecklade regioner), men även avseende möjlighet till samarbete, tillgång till teknisk kompetens eller ledningskompetens, m.m.</S></SSEG>
<TSEG ID=EN-000004.TS345><S ID=EN-000004.196.2>**The** SMEs, which make up virtually the entire economic fabric encounter special difficulties there, particularly <PHR>**with regard to**</PHR> financing (e.g. actual interest rates are often 2-3 points higher than in the more developed regions) but also <PHR>**with regard to**</PHR> cooperation opportunities, access to sources of technical or management skills, etc.</S>
</TSEG>
</SEG>

We achieve similar results by searching for Swedish phrases or parts of them, "syssel" in this case.

```
mknsg -D sgml -D sgml/pedant Swedish-English.sgm \
  | sggrep ".*/SEG" "SEG/.*/SSEG/.*/PHR" "syssel"
```

Note the change from `TSEG` to `SSEG` in the subquery of *sggrep*. A sample of the output is as follows:

<SEG ID=SE-000002.S78>
<SSEG ID=SE-000002.SS78><S ID=SE-000002.45.1 LANG=SE>Inom denna ram har Europeiska rådet uppmärksammat det italienska ordförandeskapets avsikt att inför mötet i Florens sammankalla en trepartskonferens i Rom i mitten av juni mellan regeringarna, arbetsmarknadens parter och kommissionen om <PHR>**tillväxt och sysselsättning**</PHR>.</S></SSEG>
<TSEG ID=EN-000002.TS78><S ID=EN-000002.45.1 LANG=EN><PHR>In this context</PHR>, it noted that, in preparation for the Florence meeting of <PHR>**the European Council**</PHR>, the Italian Presidency intended to hold a Tripartite Conference on <PHR>**growth and employment**</PHR>, involving governments, social partners and the Commission, in Rome in mid-June.</S></TSEG>
</SEG>
<SEG ID=SE-000003.S8>
<SSEG ID=SE-000003.SS8><S ID=SE-000003.8.1 LANG=SE>Med stöd av den strategi som det uppnåddes enighet om i Essen samt vitboken diskuterade Europeiska rådet i detalj <PHR>**tillväxt och sysselsättning**</PHR> <PHR>**på grundval av**</PHR> kommissionens meddelande "Insatser för sysselsättning i Europa: en förtroendepakt", den gemensamma interimsrapporten om sysselsättning samt de tidigare dokumenten, inklusive slutsatserna från trepartskonferensen om <PHR>**tillväxt och sysselsättning**</PHR> i Rom den 14-15 juni 1996 och Frankrikes <PHR>**memorandum om en social**</PHR> modell för Europa.</S></SSEG>

<TSEG ID=EN-000003.TS8><S ID=EN-000003.8.1 LANG=EN>Drawing on the strategy agreed in Essen and on <PHR>**the White Paper**</PHR>, <PHR>**the European Council**</PHR> held a detailed discussion on the subject of <PHR>**growth and employment**</PHR> <PHR>**on the basis**</PHR> of the Commission communication entitled "Action for employment in Europe: A confidence pact", the joint interim report on employment <PHR>**as well as**</PHR> the other documents before it, including the conclusions drawn from the Tripartite Conference on Growth and Employment held in Rome on 14 and 15 June 1996 and the French Memorandum on a European social model.</S></TSEG>
</SEG>
<SEG ID=SE-000003.S11>
<SSEG ID=SE-000003.SS11><S ID=SE-000003.9.3 LANG=SE>I överensstämmelse med kommissionens strategi gäller det att sätta igång en öppen och flexibel process som <PHR>**gör det möjligt**</PHR> för alla berörda att göra specifika åtaganden inom sina respektive ansvarsområden <PHR>**för att skapa**</PHR> en för sysselsättningen gynnsam makroekonomisk ram, maximalt utnyttja <PHR>**den inre marknaden**</PHR>s möjligheter, påskynda reformer på arbetsmarknaden och bättre utnyttja unionens politik till förmån för <PHR>**tillväxt och sysselsättning**</PHR>.</S></SSEG>
<TSEG ID=EN-000003.TS11><S ID=EN-000003.9.3 LANG=EN>In line with the Commission's approach, an open and flexible process <PHR>**needs to be**</PHR> got under way which will enable all those concerned to enter into specific commitments at their own level of responsibility <PHR>**in order to**</PHR> create a macroeconomic framework favourable to employment, to exploit to the full the potential of <PHR>**the internal market**</PHR>, to speed up <PHR>**the labour market**</PHR> reforms and to make better use of the Union's policies in the interest of <PHR>**growth and employment**</PHR>.</S></TSEG>
</SEG>

In the first translation pair, we see that "tillväxt och sysselsättning" is isolated as a multiword unit and corresponds to "growth and employment" in the English half of the translation segment. In the second segment we find the same two correspondences, the statistical processing has isolated several other multiword segments. In the English half we find "on the basis of", which corresponds to "på grundval av" in the Swedish half, and which also has been marked as significant. Similar results can be seen in the third and last translation pair. "In order to" has been isolated as a significant trigram but the infinitive falls outside the scope of the tagging. In Swedish, the translation equivalent is "för att skapa" and it has also been isolated as a significant trigram. "The internal market" corresponds to "den inre marknaden" and once again "growth and employment" corresponds to "tillväxt och sysselsättning".

Our immediate efforts will concentrate on investigating the best balance of bigrams, trigrams, quadrigrams and n-grams between the various languages. Many prepositional phrases in English, for example, are translated by fewer orthographic words in Swedish because of compounding.

We also noticed that "phrases" which end in functional words have a fairly predictable pattern of morphosyntactic tags following them and when the

equivalent phrase in another language also ends, for example, with a preposition, the same can be seen. In the previous example, for instance, "in order to" is going to be followed by an infinitive. We want to see how far this will lead us in identifying even more equivalents. In other words, if we know that we have identified phrasal equivalents between two languages and we also know the grammatical constructions that usually follow them, then we want to see if we can isolate previously unidentified equivalents — the material following the phrases — based on how well they fit the grammatical constructions that usually follow the identified phrases.

Notes

1. This is a temporary oversimplification motivated by the desire not to introduce technicalities involved with the NSL package at this early stage.
2. The most pressing need for extra annotation is in the treatment of lists where the list as a whole could be regarded as a <p> and the individual items could be equated, technically, with <s>.
3. The "document" in this case is a whole collection of texts, which make up one SGML document, a "corpus" built according to TEI's specifications. One must be careful not to mix up the terms "documents" and "files" in an SGML context.
4. The L stands for "lemmatizer" and the rest by the same analogy as PEDANT.
5. It does work, however, in some circumstances with very special texts from an industrial domain.
6. The original article appeared in *Computational Linguistics* 19: 61-74, 1993.
7. Press65, a one million word corpus of Swedish newspaper texts from 1965, has been processed by the same routines and took 12 days, nothing for someone who wants to experiment; thus the subcorpus. Over and beyond this, the $-2 \log \lambda$ is supposed to be particularly suited for small texts. The Swedish-English subcorpus that was used for this test contained 60,000+ words per language.
8. This prevents phrases such as "as soon as possible" from being tagged since "soon as possible" will be tagged first. It has been done this way in order to simplify the first experimental probes, preventing problems with SGML's inadequacy in handling overlapping structures, an inadequacy that can only be alleviated by introducing LT NSL's hyperlinking mechanism.

References

Armstrong, S. 1996. *Multilingual Corpora: Survey of Work with Multilingual Texts. Technical Report, EAGLES.* Geneva: Text Corpora Working Group, ISSCO.

Atwell, E. 1996. Machine Learning from Corpus Resources for Speech and Handwriting Recognition. Thomas, J. and M. Short (Eds.). 1996. *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*: 151–166. London/New York: Longman.

Danielsson, P. and D. Ridings. 1996a. *Annotating Parallel Texts with the NSL Library.* Research Reports from the Department of Swedish, Göteborg University GU-ISS-96-7, Språkdata.

Danielsson, P. and D. Ridings. 1996b. *PEDANT: Parallel Texts in Göteborg*. Research Reports from the Department of Swedish, Göteborg University GU-ISS-96-2, Språkdata.

Dunning, T. 1994. Accurate Methods for the Statistics of Surprise and Coincidence. Armstrong, S. (Ed.). 1994. *Using Large Corpora*: 61–74. Cambridge, MA./London: The MIT Press.

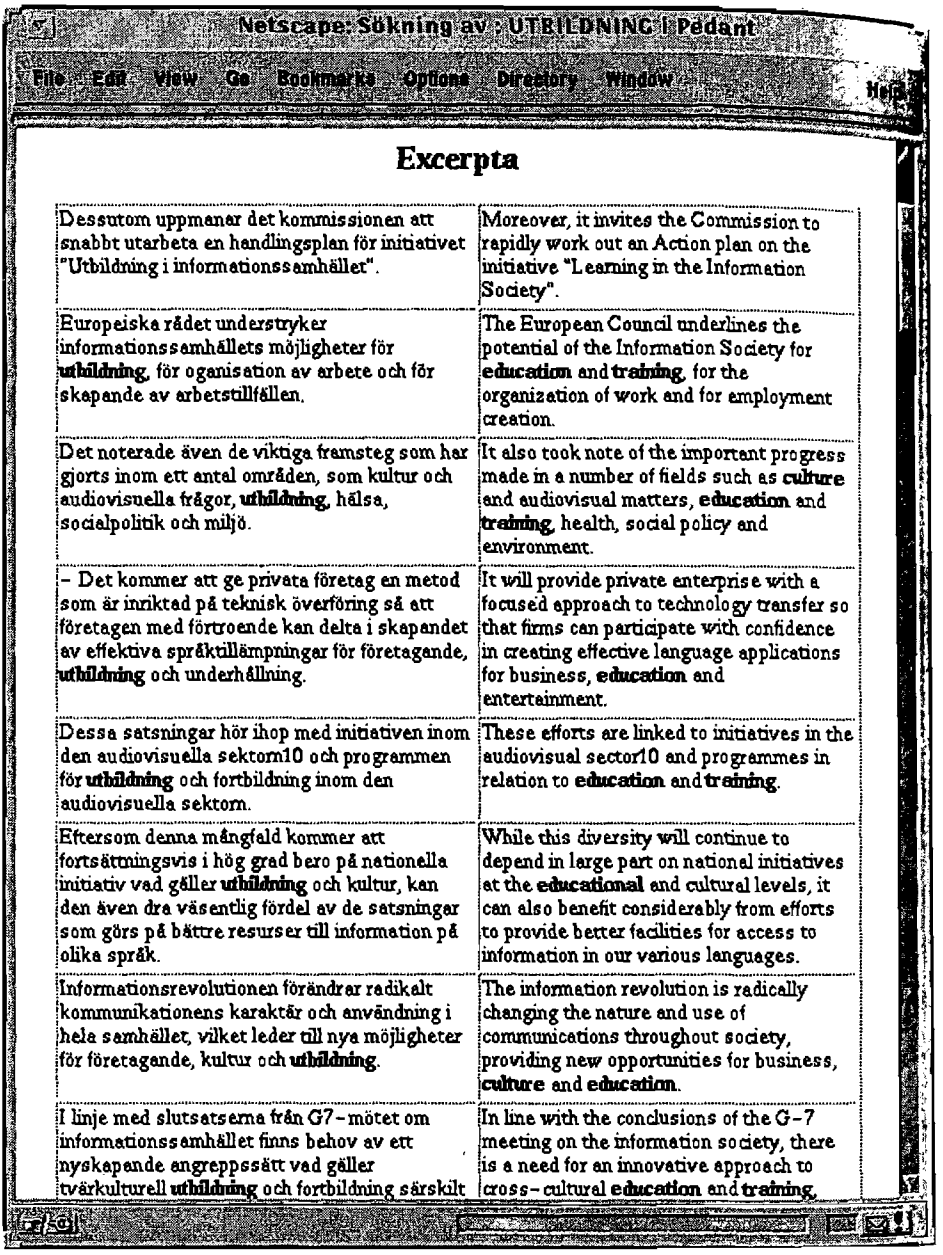
Hellberg, S. 1978. *The Morphology of Present-day Swedish: Word-inflection, Word-formation, Basic Dictionary*. Data linguistica. Stockholm: Almqvist & Wiksell International.

Karlsson, F. 1992. SWETWOL: A Comprehensive Morphological Analyser for Swedish. *Nordic Journal of Linguistics* 15: 1–45.

Sperberg-McQueen, C.M. and L. Burnard (Eds.). 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago: ACH, ACL, ALLC.

Thompson, H., S. Finch and D. McKelvie. 1995. *Normalised SGML Library (NSL)*. Multext LRE Project 62-050, University of Edinburgh, The Language Technology Group.

Appendix



Dessutom uppmanar det kommissionen att snabbt utarbeta en handlingsplan för initiativet "Utbildning i informationssamhället".	Moreover, it invites the Commission to rapidly work out an Action plan on the initiative "Learning in the Information Society".
Europeiska rådet understryker informationssamhällets möjligheter för **utbildning**, för oganisation av arbete och för skapande av arbetstillfällen.	The European Council underlines the potential of the Information Society for **education** and **training**, for the organization of work and for employment creation.
Det noterade även de viktiga framsteg som har gjorts inom ett antal områden, som kultur och audiovisuella frågor, **utbildning**, hälsa, socialpolitik och miljö.	It also took note of the important progress made in a number of fields such as **culture** and audiovisual matters, **education** and **training**, health, social policy and environment.
– Det kommer att ge privata företag en metod som är inriktad på teknisk överföring så att företagen med förtroende kan delta i skapandet av effektiva språktillämpningar för företagande, **utbildning** och underhållning.	It will provide private enterprise with a focused approach to technology transfer so that firms can participate with confidence in creating effective language applications for business, **education** and entertainment.
Dessa satsningar hör ihop med initiativen inom den audiovisuella sektorn10 och programmen för **utbildning** och fortbildning inom den audiovisuella sektorn.	These efforts are linked to initiatives in the audiovisual sector10 and programmes in relation to **education** and **training**.
Eftersom denna mångfald kommer att fortsättningsvis i hög grad bero på nationella initiativ vad gäller **utbildning** och kultur, kan den även dra väsentlig fördel av de satsningar som görs på bättre resurser till information på olika språk.	While this diversity will continue to depend in large part on national initiatives at the **educational** and cultural levels, it can also benefit considerably from efforts to provide better facilities for access to information in our various languages.
Informationsrevolutionen förändrar radikalt kommunikationens karaktär och användning i hela samhället, vilket leder till nya möjligheter för företagande, kultur och **utbildning**.	The information revolution is radically changing the nature and use of communications throughout society, providing new opportunities for business, **culture** and **education**.
I linje med slutsatserna från G7-mötet om informationssamhället finns behov av ett nyskapande angreppssätt vad gäller tvärkulturell **utbildning** och fortbildning särskilt	In line with the conclusions of the G-7 meeting on the information society, there is a need for an innovative approach to cross-cultural **education** and **training**,

Figure 1: The Netscape view of the database

A B	A ~B	~A B	~A ~B	log λ	trigram
114	2	6	62954	1599.97	små och medelstora
89	31	163	62793	881.92	och medelstora företag
59	4	76	62937	726.15	när det gäller
40	61	1	62974	524.01	informations- och kommunikationsteknologi
40	61	8	62967	490.18	informations- och kommunikationsteknologin
42	0	1343	61691	322.02	inom ramen för
19	0	57	63000	260.61	snart som möjligt
20	0	141	62915	241.42	mellan Europeiska unionen
23	22	80	62951	238.45	forskning och utveckling
12	3	1	63060	207.55	den 1 januari
17	6	59	62994	206.25	gör det möjligt
16	9	57	62994	187.51	på nationell nivå
20	0	582	62474	186.74	i enlighet med
13	3	26	63034	181.57	den inre marknaden
10	1	2	63063	177.47	den tredje etappen
13	10	15	63038	176.50	på detta område
10	7	0	63059	171.95	hälso- och sjukvården
9	0	2	63065	166.96	formerna för arbetets
18	2	431	62625	165.77	i fråga om
11	0	30	63035	164.70	noterar med tillfredsställelse
13	7	42	63014	160.62	tillväxt och sysselsättning
11	0	43	63022	157.80	varor och tjänster
25	0	2678	60373	157.72	av informations- och
21	234	40	62781	154.98	Europeiska rådet uppmanar
11	21	4	63040	153.81	den gemensamma valutan
19	236	23	62798	153.19	Europeiska rådet välkomnar
11	12	9	63044	153.03	I detta sammanhang
11	0	62	63003	150.50	på hög nivå
9	3	3	63061	150.40	1 januari 1999
9	0	11	63056	149.86	för arbetets organisation
10	21	2	63043	145.19	de nya formerna
11	1	62	63002	143.62	på europeisk nivå
8	1	2	63065	143.27	Central- och Östeuropa
19	2	1146	61909	138.85	så snart som
25	9	1844	61198	137.51	det möjligt att
16	1	734	62325	134.58	med hänsyn till
13	6	148	62909	132.64	i Europeiska unionen
11	5	65	62995	129.68	göra det möjligt
16	239	19	62802	129.25	Europeiska rådet noterar
20	0	2683	60373	126.14	för små och
7	2	1	63066	125.92	äldre och handikappade
9	1	41	63025	123.75	att lägga fram
8	1	17	63050	121.94	Den tredje utmaningen
21	359	64	62632	121.56	för att skapa
13	0	589	62474	121.23	i samband med
8	1	19	63048	120.47	experter på hög
9	0	80	62987	119.08	på så sätt
9	17	9	63041	118.90	frågor som rör
9	4	32	63031	118.19	uttrycker sin tillfredsställelse
6	0	1	63069	117.38	Frågor som oroar
7	0	12	63057	116.48	på lång sikt
13	0	737	62326	115.46	anpassa *sig till*
6	0	2	63068	114.13	Frågor att fundera

Figure 2: Swedish trigrams

A B	A ~B	~A B	~A ~B	log λ	trigram
180	39	171	66968	1801.10	The European Council
57	2	27	67272	797.51	the Information Society
86	146	101	67025	754.48	the European Union
51	4	114	67189	602.21	the Member States
47	6	86	67219	566.93	education and training
61	2	1862	65433	418.12	in order to
24	3	3	67328	391.43	Route of Actions
55	177	296	66830	334.77	the European Council
23	18	21	67296	296.06	with a view
29	2	323	67004	292.37	as well as
50	5	2855	64448	282.11	the development of
23	209	2	67124	249.29	the European Parliament
18	2	42	67296	245.88	soon as possible
18	15	18	67307	236.81	the social partners
24	27	141	67166	221.82	the United States
34	0	2871	64453	214.15	in terms of
20	0	332	67006	211.32	as soon as
22	24	129	67183	208.20	research and development
15	36	5	67302	198.01	the United Kingdom
13	2	18	67325	194.43	notes with satisfaction
16	4	81	67257	192.16	the labour market
21	13	277	67047	184.05	to ensure that
10	0	3	67345	182.26	the Structural Funds
22	457	13	66866	172.64	in the field
14	18	23	67303	172.47	In this context
12	0	50	67296	170.26	the Intergovernmental Conference
20	227	22	67089	168.01	European Council welcomes
62	749	533	66014	167.09	of the European
9	0	3	67346	165.07	Central and Eastern
23	0	1900	65435	163.85	a view to
11	0	37	67310	162.16	this Green Paper
14	5	89	67250	161.63	at Community level
16	41	50	67251	158.32	of new technologies
26	193	161	66978	151.22	The European Union
19	15	370	66954	150.30	have to be
9	0	21	67328	141.92	Council took note
10	41	1	67306	139.12	the United Nations
8	3	2	67345	137.71	1 January 1999
9	1	18	67330	137.70	Competitiveness and Employment
9	1	19	67329	136.90	Heads of State
23	1	2882	64452	136.55	the field of
9	7	8	67334	133.13	the single currency
9	0	39	67310	132.24	the White Paper
21	0	2884	64453	132.18	the end of
9	9	7	67333	131.69	Paper on Growth
22	1	2883	64452	130.34	the use of
13	11	124	67210	129.35	growth and employment
15	192	19	67132	128.11	on the basis
24	0	4676	62658	127.91	note of the
11	4	86	67257	127.86	the internal market
19	1	1904	65434	127.43	with regard to

Figure 3: English trigrams