Published in: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, pages 531–538, Athens, Greece, May/June 2000.

A Parallel Corpus of Italian/German Legal Texts

Johann Gamper

European Academy Bolzano
Scientific Area "Language and Law"
Weggensteinstr. 12a, 39100 Bozen, Italy
jgamper@eurac.edu

Abstract

This paper presents the creation of a parallel corpus of Italian and German legal documents which are translations of one another. The corpus, which contains approximately 5 mio. words, is primarily intended as a resource for (semi-)automatic terminology acquisition. The guidelines of the Corpus Encoding Standard have been applied for encoding structural information, segmentation information, and sentence alignment. Since the parallel texts have a one-to-one correspondence on the sentence level, building a perfect sentence alignment is rather straightforward. As a result of this the corpus constitutes also a valuable testbed for the evaluation of alignment algorithms. The paper discusses the intended use of the corpus, the various phases of corpus compilation, and basic statistics.

1. Introduction

Electronic text corpora are valuable resources in all areas dealing with natural language processing in one form or another: Tools such as part-of-speech taggers or morphological analyzers are trained, language models for machine translation are inferred, lexical and terminological data are acquired, etc. Starting with the resurgence of interest in empirical and statistical methods in natural language processing in the 1990s and the increasing availability of language material in electronic form, much effort has been spent in building and investigating large corpora for a variety of languages, subject fields, and applications (Rubio et al., 1998).

In the field of terminology, large collections of domain-specific language material are explored for terminologically relevant information. While the manual acquisition is a time-consuming and error-prone process, recent advances in corpus-based research brought computer programs which scan the corpus for terminological data and generate lists of term candidates which have to be post-edited by humans (Dagan and Church, 1997; Heid et al., 1996). Such a computer-assisted approach has several advantages over manual term acquisition and improves terminological research and its output by opening new doors for empirical investigation such as an increased efficiency and scope of work or exhaustive search for terms (Bowker, 1996).

Before a corpus is useful for automatic exploration by computer programs, the interpretation of various chunks of the text have to be made explicit. This process is known as encoding. Marking up is a widely used technique for corpus encoding. Tags or markup are interspersed with the original text and represent the interpretation of the enclosed text segments. The increased interest in corpus-based research favored the development of standards and languages for text encoding and interchange such as the TEI guidelines (Sperberg-McQueen and Burnard, 1994), the Corpus Encoding Standard (Ide et al., 1996; Ide, 1998), SGML (Goldfarb, 1990), and recently XML (Bray et al., 1998).

In this paper we present the compilation of a parallel corpus of Italian/German legal texts. The corpus is part of a larger project about computer-assisted terminology acquisition, which is briefly introduced in section 2. The sections 3.–5. describe the corpus compilation process in detail. Basic corpus statistics are presented in section 6.

2. Background

Due to the equal status of the Italian and German languages in South Tyrol, a great part of legal and administrative documents has to be translated into the other language, respectively. This requires an independent German legal language for South Tyrol, which reflects the Italian legislation and hence is different from the legal language in other German-speaking countries. The basis for such a language is a consistent and comprehensive terminology, which, however, does not exist yet. While organized terminological activities have been neglected for decades, various institutions coined different German terms with the effect that one and the same Italian term has been translated into different German terms and vice versa. The result of this is a "terminological chaos", lots of inconsistencies, duplication of efforts, and poor quality translations. A consistent terminology forming the basis for a German legal language in South Tyrol would substantially contribute to solve these problems.

Since 1994 the European Academy Bolzano, in cooperation with the Joint Terminology Committee, has been working on a standardization of the legal terminology in South Tyrol (Arntz and Mayer, 1996). Italian/German term pairs are collected from parallel text material. In particular cases further terminological investigation is required such as a comparison with the terminology in the Austrian/German/Swiss law systems. As a result, a first Italian/German dictionary of legal terms has been published (Mayer, 1998).

Experience gathered at the European Academy Bolzano has shown that exhaustive analyses of a huge number of relevant texts require the use of advanced computational methods. The CATEX (Computer Assisted Terminology Extraction) project emerged from this need to support and improve, both qualitatively and quantitatively, the manual acquisition of terminological data. The main objective of CATEX is to develop a computational framework for (semi-)automatic terminology acquisition which con-

sists of the following modules: a parallel text corpus, term-extraction programs, and a terminology database with links to the corpus.

In the rest of the paper we discuss the creation of the parallel text corpus, which comprises the following steps:

- corpus design,
- pre-processing,
- encoding primary data, and
- encoding linguistic annotation.

A graphical overview is shown in figure 2.

While the corpus is mainly intended as a resource for terminological investigation, it could serve for other applications as well. Due to the one-to-one correspondence of the texts on the sentence level, the corpus can serve as valuable testbed for the evaluation of alignment algorithms. Finally, the large amount of text material allows linguistic analyses of the Italian and German legal languages.

3. Corpus Design

Corpus design selects a collection of texts which should be included in the corpus. An important selection criteria is that the texts represent a realistic model of the language to be studied. Bowker (Bowker, 1996) mentions three text types which should be included in a special language corpus: instructional, advanced, and popularized texts. Other criteria are text size, corpus size, authors, publication date, etc.

Our corpus consists of Italian and German legal documents, which are translations of one another. In its current form, the corpus contains approx. 5 mio. words. Only one sort of texts is included: the bilingual version of important Italian laws including legislative decrees, etc. (see table 1). The national laws are originally written down in Italian. Starting in 1982, the most important law books have been translated into German. The provincial laws are originally written down in German and then translated into Italian. The translation of provincial laws started in the 1950s.

While this collection of texts is certainly not representative for the legal language in everyday use, it builds the core part for the compilation of a legal terminology in South Tyrol. The books are accepted as standard translations and used as "dictionaries" by translators, hence the corresponding terminology is widely used. At a later phase we plan to extend the corpus with additional text material such as sentences, administrative documents, judgments, etc.

A particular feature of our corpus is the structural equivalence of the original text and its translation down to the sentence level, i.e. each sentence in the original text has a corresponding one in the translation.

4. Pre-Processing

A great part of the texts had to be OCRed from a printed version. Other law books were already available in electronic form in different formats. In the pre-processing phase we correct (mainly OCR) errors in the raw text material and produce a unified electronic version in order to simplify the programs for consequent encoding. An excerpt from the Civil Code is given in figure 1 and will be used as an example in the rest of the paper.

The first step is to specify a character set (including letters, numbers, punctuation symbols, etc.) for both languages and to map the texts into these sets. Especially when small fonts are used, the German letters "ä", "ö", etc. and the Italian letters "à", "è", etc. are often recognized as characters outside the specified ranges. In other cases the OCR program makes mismatches between similar characters, e.g. "I", "l", and "1". This yields tokens consisting of letters and digits which easily can be detected and removed.

Another frequent type of OCR error concerns the correct recognition of paragraph breaks. The OCR-program recognizes the end of a paragraph if there is a space between the end of the last line and the right border. In cases where the last sentence coincides with the right border, the end of a paragraph cannot be detected. On the other hand, titles which are centered and longer than one line are divided into more than one paragraph, since usually there is some space between the end of each line and the right border.

Another step in the pre-processing phase is some kind of standardization. While the Italian original titles are numbered using Roman numbers, most German translations use Arabic numbers. We replace these Arabic numbers by Roman numbers, e.g. "`1. Titel`" is translated into "`Titel I`".

The following lines show the German part of our example text after the pre-processing step:

```
<6>Buch I
<5>Personen- und Familienrecht
<6>Titel I
<5>Natürliche Personen
<3>1.  (Rechtsfähigkeit)
<1>Die Rechtsfähigkeit wird zum
Zeitpunkt der Geburt erworben (22
Verf.).
Die Rechte, die das Gesetz dem
Gezeugten zuerkennt, hängen von der
tatsächlichen Geburt ab (254, 462,
784).<8>1<1>)
<8>1<4>) Der dritte Absatz wurde
durch Artikel 1 des Königlichen
Gesetzesdekrets vom 20.1.1944, Nr.
25, und durch Artikel 3 der
gesetzesvertretenden Verordnung des
Statthalters vom 14.9.1944, Nr. 287,
aufgehoben.
```

The tags indicate formatting information such as various font types and sizes: <1> is normal size, <3> is normal size boldface, <4> is small, <6> is large, <5> is large boldface, and <8> is superscript.

5. Corpus Encoding

This section discusses in detail corpus encoding — a process which adds various pieces of information to the raw text material. We apply the Corpus Encoding Standard, which roughly distinguishes between primary data (which is the raw text material) and linguistic annotation (which is information resulting from linguistic analyses of texts).

Law Books	Size (in kW) Italian	German
National laws		
– Codice Civile	244	255
– Codice di Procedura Civile	108	115
– Leggi Complementari al Codice Civile	111	116
– Codice di Procedura Penale	133	140
– Ordinamento del Notariato Italiano	49	51
– Fallimento ed altre Procedure Concorsuali	32	32
– Testo unico delle Imposte sui Redditi	46	44
– Processo Amministrativo	26	26
Provincial laws		
– Codice della Provincia Autonoma di Bolzano	1,791	1,623
Total	2,540	2,402

Table 1: Law books included in the corpus.

Libro I.
Delle persone e della famiglia

Titolo I.
Delle persone fisiche.

1. (Capacità giuridica).

La capacità giuridica si acquista dal momento della nascita (22 Cost.).

I diritti che la legge riconosce a favore del concepito sono subordinati all'evento della nascita (254, 462, 784).[1])

[1]) Il comma 3 è stato abrogato in virtù dell'art. 1 R.D.L. 20 gennaio 1944, n. 25 e dell'art. 3 D.Lg.Lt. 14 settembre 1944, n. 287.

1. Buch
Personen- und Familienrecht

1. Titel
Natürliche Personen

1. (Rechtsfähigkeit)

Die Rechtsfähigkeit wird zum Zeitpunkt der Geburt erworben (22 Verf.).

Die Rechte, die das Gesetz dem Gezeugten zuerkennt, hängen von der tatsächlichen Geburt ab (254, 462, 784).[1])

[1]) Der dritte Absatz wurde durch Artikel 1 des Königlichen Gesetzesdekrets vom 20. 1. 1944, Nr. 25, und durch Artikel 3 der gesetzesvertretenden Verordnung des Statthalters vom 14. 9. 1944, Nr. 287, aufgehoben

Figure 1: Civil Code excerpt: Italian original and German translation.

5.1. Encoded Information

Since corpus encoding is very costly, a careful analysis about what information should be added to the text is recommended. For the purpose of computer-assisted terminology extraction the following pieces of information are required: bibliographic information, structural information, segmentation of the texts into sentences and tokens, lemmas, part-of-speech tags, and alignment.

Bibliographic and structural information is required to automatically compute the document and the part inside the document where a term has been found, e.g. "Codice civile, art. 12". Moreover, structural information facilitates a user-friendly navigation through the corpus. This is important in our case, since the corpus will be linked to the terminology database to provide a rich source of contextual information.

The automatic extraction of terminological units and their translation equivalents is mainly inspired by the work in (Dagan and Church, 1997). The monolingual identification of terms is based on part-of-speech patterns which characterize valid terms. The recognition of the corresponding translation equivalents requires parallel texts which are aligned on the word level. Lemmas abstract from singular/plural variations, which is useful for both alignment and term recognition.

5.2. Corpus Encoding Standard

The Corpus Encoding Standard (CES) has been developed to provide a set of guidelines for encoding corpora for language engineering applications (Ide et al., 1996; Ide, 1998). CES is an application of SGML (Goldfarb, 1990), a declarative markup language which has been designed to be a standard to describe the content and structure of a document independently of any formatting information. CES is conformant to the TEI Guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative (Sperberg-McQueen and Burnard, 1994). While the TEI Guidelines are designed towards maximum applicability across a broad range of applications, CES is a subset thereof, optimally suited for applications in natural language engineering.

CES distinguishes primary data, which is the raw text material in electronic form, and linguistic annotation, which is information resulting from linguistic analyses of raw texts. Primary data encoding covers the markup of relevant objects, such as the structure, in the raw text material. SGML tags are interspersed with the primary data to make the interpretation of text chunks explicit. For the linguistic annotation, CES recommends to store this information in separate SGML files which are linked to the original or to other annotation documents. This leads to an organization of the text corpus as shown in figure 2.

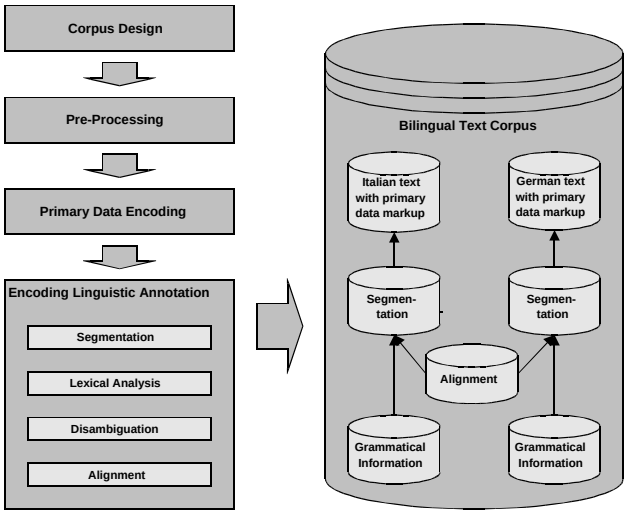

Figure 2: Corpus compilation and organization of the corpus files.

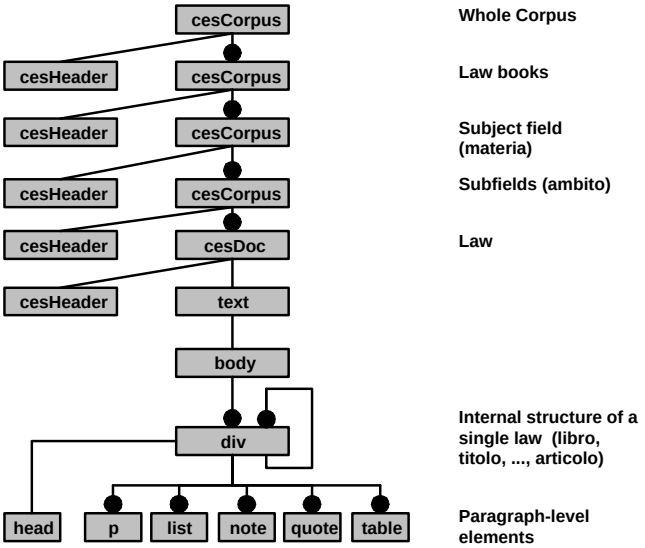

Figure 3: Upper level of the tree structure of the `cesDoc` file for the provincial law book.

As an SMGL application, the structure of document classes are defined in so-called document type definitions (DTD). CES provides three different types of DTDs: `cesDoc` for primary data encoding, `cesAna` for the encoding of segmentation and grammatical information, and `cesAlign` for the alignment of parallel texts. Recently, an XML version of these DTDs has been provided too.

5.3. Encoding Primary Data

Primary data encoding covers the markup of relevant objects in the raw text material comprising documentation and structural information. Documentation information includes global information about the text such as title, author(s), used character sets, encoding conventions, etc. Regarding the text structure, CES distinguishes between gross structural markup (structural elements down to the paragraph level, e.g. sections, lists, paragraphs, footnotes, etc.) and markup for sub-paragraph elements (abbreviations, references, dates, names, etc.).

The law books show a quite pronounced hierarchical structure up to 9 levels. A distinction between the grouping and the internal structure of laws is useful. Each book contains a collection of laws which might be grouped into subject fields, subfields, etc. For example, the provincial law book contains 929 laws (ordinary laws, legislative decrees, etc.) grouped into 38 main subject fields (materia), each of which is further divided into several subfields (ambito). The internal structure of a single law shows hierarchically ordered divisions up to 7 levels: parte, libro, titolo, capo, sezione, paragrafo, articolo.

For primary data encoding the `cesDoc` DTD is used. Figure 3 shows the upper level of the `cesDoc` file for the provincial law book. The structural units of the law book listed on the right-hand side are encoded as SGML elements as shown in the tree structure on the left-hand side.

We made extensive use of the recursive feature of the <`cesCorpus`> element which allows us to divide the corpus into several subcorpora. The entire corpus is encoded as a <`cesCorpus`> element containing a <`cesHeader`> and one <`cesCorpus`> element for each law book. The <`cesHeader`> elements contain global information about the corresponding (sub)corpus or document. The grouping of laws into subject fields and subfields is encoded in form of nested <`cesCorpus`> elements. A law is considered as a single document and encoded as a <`cesDoc`> element, which contains a <`text`> and a <`body`> element. The internal structure of laws is represented in form of nested <`div`> elements. The `type` attribute holds the type of the division, e.g. libro, titolo, etc. The smallest division is of type `articolo`. On the paragraph level we encoded the following objects: paragraphs (<`p`>), lists (<`list`>), tables (<`table`>), quotes (<`quote`>), and footnotes (<`note`>). Each element has a unique identifier which is stored in the `id` attribute and serves as a reference system for various purposes such as references within the corpus or alignment.

On the sub-paragraph level, we encoded the following objects: dates (<`date`>), abbreviations (<`abbr`>), numbers (<`num`>), references to parts in the same law or in other laws (<`ref`>), footnote references (<`ptr`>), and sentences (<`s`>). The reference to and the number of footnotes have been removed from the primary data and encoded in the `n` attribute, since these parts of the text are not relevant to corpus analyses. Similar, the labels of lists are considered as rendition information and are encoded in the `n` attribute of <`item`> elements.

A simplified version of the result of encoding structural information is given below:

```
<body id="cc">
 <p id="cc.0.0.0.p1">
  Zivilgesetzbuch
 </p>
 <div type="libro" id="cc.1.0.0" n="I">
  <head id="cc.1.0.0.h1">
   Personen- und Familienrecht
  </head>
  <div type="titolo" id="cc.1.1.0" n="I">
   <head id="cc.1.1.0.h1">
```

```
 Natürliche Personen
</head>
<div type="articolo" id="cc.1.1.1">
 <head id="cc.1.1.1.h1">
  1. (Rechtsfähigkeit)
 </head>
 <p id="cc.1.1.1.p1">
  Die Rechtsfähigkeit wird zum Zeitpunkt
  der Geburt erworben
  (22 <abbr>Verf.</abbr>).
 </p>
 <p id="cc.1.1.1.p2">
  Die Rechte, die das Gesetz dem
  Gezeugten zuerkennt, hängen von der
  tatsächlichen Geburt ab (254, 462, 784).
  <ptr target="cc.1.1.1.fn1" n="1">
 </p>
 <note id="cc.1.1.1.fn1" n="1">
  Der dritte Absatz wurde durch Artikel
  1 des Königlichen Gesetzesdekrets vom
  <date>20.1.1944</date>,
  <abbr>Nr.</abbr> 25, und durch
  Artikel 3 der gesetzesvertretenden
  Verordnung des Statthalters vom
  <date>14.9.1944</date>,
  <abbr>Nr.</abbr> 287, aufgehoben.
 </note>
</div>
```

Encoding structural information amounts to (1) translating the presentation information in the raw text material into structural elements, (2) eliminating presentational markup which does not point to a relevant object, and (3) adding markup for relevant objects not marked in any way in the raw text material. For example, "<6>Buch I" (which means that "Buch I" is printed in a big font) marks the beginning of a structural unit and is translated into — a division of type libro, where the attribute n stores the number of the unit. The same tag <6> followed by "Titel" is translated into a division of type titolo.

5.4. Encoding Linguistic Annotation

Encoding linguistic annotation enriches the primary data with information resulting from linguistic analyses of the raw text material. We consider the following steps:

- tokenization and segmentation,
- the assignment and disambiguation of lemmas and POS tags,
- sentence and word alignment.

Following the CES guidelines, the information resulting from these analyses is stored in separate SGML files of type cesAna (see figure 2).

5.4.1. Tokenization and Segmentation

Tokenization is the task of splitting the input text into a sequence of tokens and to assign to each of these tokens a label indicating its type such as abbreviation, punctuation, etc. (Habert et al., 1998). This seemingly trivial task can be difficult. Words may contain other characters than letters such as in "AT&T". Detecting multi-word units such as "Rechts- und Handlungsfähigkeit" is difficult. The isolation of punctuation symbols, and hence the identification of sentence boundaries, is not straightforward, since various punctuation symbols might be used otherwise as well, e.g. periods in abbreviations. In a sentence ending with "etc.", the period both is part of the abbreviation and indicates the end of the sentence.

The identification and classification of punctuation symbols yields a segmentation of the input text into segments such as sentences. This process is also known as segmentation (Armstrong, 1996). There is no universal definition of what constitutes a sentence. For our purposes we follow the definition in (Habert et al., 1998) and, in addition to text segments terminated by a full-stop punctuation symbol, consider the following segments as sentences: titles, items of an enumeration, and table cells.

An example of a cesAna document which stores the result of tokenization and segmentation is given below:

```
<chunk doc="cc.de.pd3">
  <par from="cc.1.1.1.p1">
    <s from="cc.1.1.1.p1.s1">
      <tok class='TOK'>
        <orth>Die</orth>
      </tok>
      <tok class="TOK">
        <orth>Rechtsfähigkeit</orth>
      </tok>
      <tok class="TOK">
        <orth>wird</orth>
      </tok>
      <tok class="TOK">
        <orth>zum</orth>
      </tok>
      <tok class="TOK">
        <orth>Zeitpunkt</orth>
      </tok>
      <tok class="TOK">
        <orth>der</orth>
      </tok>
      <tok class="TOK">
        <orth>Geburt</orth>
      </tok>
      <tok class="TOK">
        <orth>erworben</orth>
      </tok>
      <tok class="OPUNCT">
        <orth>(</orth>
      </tok>
      <tok class="DIG">
        <orth>22</orth>
      </tok>
      <tok class="ABBR">
        <orth>Verf.</orth>
      </tok>
      <tok class="CPUNCT">
        <orth>)</orth>
      </tok>
      <tok class="PTERM_P">
        <orth>.</orth>
      </tok>
    </s>
```

The `class` attribute of the <tok> element specifies the class of a token. The `from` attribute in the <par> and <s> elements are links to the `cesDoc` file which contains the primary data markup.

We used the MULTEXT[1] tokenizer `MtSeg` which splits the input text into a sequence of tokens and detects sentence boundaries. Eleven classes of tokens are distinguished: abbreviations, dates, numbers, enumerations, various punctuation types, etc. In a first step, tokens are identified and labeled with the most general class `TOK`. In further refinement steps, the initial assignment might be revised when tokens are recognized to belong to more specific classes such as abbreviation, digit, etc.

5.4.2. Alignment

Alignment of parallel texts can be defined as the task of identifying corresponding parts between a text and its translation. The alignment can be done between parts on various levels of granularity, e.g. between divisions, paragraphs, sentences, words, phrases, characters. Aligned text corpora have proved to be very useful in a number of applications such as bilingual lexicography and machine translation (Simard, 1998). Several programs for text alignment have been developed in the past, e.g. (Brown et al., 1991; Gale and Church, 1994) for sentence alignment, (Dagan et al., 1993; Melamed, 1997) for word alignment, and (Church, 1993) for character alignment.

While sentence alignment is not that difficult, many applications such as our approach to automatic terminology acquisition require word alignment, which turns out to be a much more complex task which still needs a lot of research. However, sentence alignment is a useful pre-processing step which improves word alignment and already provides useful information for browsing parallel corpora.

Provided that the segmentation into sentences is correct, the alignment of our corpus on the sentence level turns out to be particularly simple. There is a rule for legal translation which says that legal documents have to be translated literally sentence by sentence (as far as possible). Hence, our corpus has the following properties:

- the original text and its translation have the same structure down to the sentence level,
- there are no ommissions or insertions of sentences, i.e. sentences which haven't been translated,
- the order of the sentences is the same.

In very few cases the translation rule has been violated, i.e. one sentence has been translated into more than one sentence or vice versa. These cases have been validated by hand and the correct alignment has been established.

The sentence alignment is stored in a `cesAlign` file as shown in the following excerpt:

```
<linkgrp targType="s" fromDoc="cc.it.ana"
                      toDoc="cc.de.ana">
  <link xtargets="cc.1.0.0.h1.s1;
                  cc.1.0.0.h1.s1">
  <link xtargets="cc.1.1.0.h1.s1;
                  cc.1.1.0.h1.s1">
  <link xtargets="cc.1.1.1.p1.s1;
                  cc.1.1.1.p1.s1">
  <link xtargets="cc.1.1.1.p2.s1;
                  cc.1.1.1.p2.s1">
  <link xtargets="cc.1.1.1.p3.s1;
                  cc.1.1.1.p3.s1">
```

5.4.3. Implementational Issues

The general approach we adopt for primary data encoding is to pass the raw texts through a sequence of filters. Each filter incrementally adds small pieces of new information and writes a logfile in case of doubt. The output and the logfile are analyzed and used to improve the filter programs in order to minimize manual post-editing. This modular bootstrapping approach has advantages over huge parametrizable programs: filters are relatively simple; tuning the filters becomes less complex; when recovering from a previous stage the loss of work is minimized. The filters are implemented in Perl which, due to its pattern matching mechanism via regular expressions, is a very powerful language for such applications. Moreover, the basic utilities in a UNIX shell and an editor which supports regular expressions such as `emacs` are very helpful.

Primary data encoding might be very costly, depending on many factors as for example how much formatting information is present in the raw text material and how well it translates into structural elements, how much information should be encoded, and how accurate the result should be. The pre-processing phase proved to be very useful to reduce the complexity of the programs for primary data encoding. However, our experience tells us that, even after pre-processing and after improving the filter programs on a really satisfactory level, the compilation of a very "clean" corpus still requires a lot of manual post-editing. Hence, the tradeoff between high accuracy and required manual work should be analyzed carefully.

For tokenization and segmentation we used the MULTEXT tokenizer `MtSeg`. The tokenizer can be customized via language-specific resource files which specify such things as abbreviations, rules determining how to treat punctuation symbols, compound words, various date formats, etc. Unfortunately, the available version of `MtSeg` cannot read SGML documents on its input, which proves to be very inconvenient for its use. An evaluation of 10% of the Civil Code ($\approx$ 28,000 words) revealed only one type of tokenization error in German: a full stop that is not part of an abbreviation and is followed by an uppercase letter is recognized as end-of-sentence marker, e.g. in "6. Absatz". This kind of error is unavoidable in German if we refuse to mark such patterns as compounds.

6. Corpus Statistics

The corpus consists of 82 national laws (originally written in Italian and then translated into German) and 929 provincial laws (originally written in German and then translated into Italian). As shown in table 1, in both collections the translation is sligthly longer in terms of words as the original version.

[1] The MULTEXT tools are available from http://www.lpl.univ-aix.fr/projects/multext.

	Italian	German
Tokens	541,946	577,095
Types	14,469	19,654
Type/token ratio	2.67	3.4
Sentences	35,752	35,732
Sentence alignments (it–de)	35,723	
1-1	35,687	
2-1	27	
1-2	7	
3-1	1	
1-3	1	
Avg. sentence length	15.15	16.14
title	2.9	2.89
paragraph	25.98	27.16
footnote	9.02	13.83
Longest sentence		
paragraph	187	204
footnote	190	238
Avg. word length	5.54	6.37
1-letter words	39,042	475
2-letter words	91,269	30,333
3-letter words	67,475	179,247
4-letter words	31,428	52,597
5-letter words	66,049	44,362
6-letter words	41,002	44,636
7-letter words	46,677	37,771
8-letter words	49,599	32,679
9-letter words	36,499	32,867
10-letter words	26,774	26,007
11-letter words	18,000	25,244
12-letter words	13,639	23,013
13-letter words	8,159	12,640
>13-letter words	6,333	35,224

Table 2: Basic statistics over a part of the national laws.

Basic corpus statistics have been performed over the following law books: Codice Civile, Codice di Procedura Civile, Leggi Complementari al Codice Civile, and Codice di Procedura Penale. Together they make up approximately 80% of the national laws and 25% of the whole corpus. Table 2 summarizes the results.

The count of the tokens does not include the following classes: dates, abbreviations, digits, enumerations, and punctuation symbols. The German part, which is the translation, is about 6.5% longer than the Italian part. The number of types and the type/token ratio should be considered carefully, since neither of the texts has been lemmatized.

An important feature of our corpus is the structural equivalence of the original text and the translation down to sentence level, which is confirmed by the numbers about sentences and sentence alignments. The Italian text is only 20 sentences longer than the German text and only 36 out of 35,723 alignments do not represent a ont-to-one correspondence. The overall corpus is expected to contain approximately 100,000 one-to-one sentence alignments. This reprents a rather large corpus with perfect sentence alignment, which could serve as a very useful reference tool for the evaluation of alignment algorithms.

On average, German sentences tend to be slightly longer than Italian sentences. To get a more precise picture of the average sentence length in the legal domain a distinction in three classes seems useful: sentences in titles, sentences in ordinary paragraphs (text which makes up the articles in a law), and sentences in footnotes. The average sentence length in ordinary paragraphs confirms the hypothesis that legal texts contain rather long sentences. A detailed analysis revealed that the average sentence length in the book "Leggi Complementari al Codice Civile" is even longer: 29 words for Italian and 30 words for German. If we include dates, digits, and abbreviations in the statistics, the average sentence length increases of about 1.5 words.

Similar as in the case of sentences, the average length of words is sligthly higher in German than in Italian. Since we didn't eliminate stopwords, this is clearly not a realistic picture of the average word length in legal language.

7. Conclusion

The need for large corpora in many applications in the field of natural language processing brought an increasing number of corpus projects over the last years (see for example (Rubio et al., 1998)). In this paper we presented first results we have achieved in encoding an Italian/German parallel corpus of legal documents. The corpus is part of an ongoing research project on computer-assisted terminology acquisition. The overall approach of corpus compilation comprises four consecutive steps: corpus design, preprocessing, primary data encoding, and encoding linguistic annotation. In its current version, the corpus contains only one type of texts, namely Italian laws. A particular characteristic of the included texts is the structural equivalence down to sentence level between the original version and the translation. This simplifies a perfect sentence alignment and makes the corpus a valuable testbed for the evaluation of alignment algorithms. The Corpus Encoding Standard has been applied for encoding structural information, segmentation information, and sentence alignment.

Future work will include the completion of the linguistic annotation including lemmatization, part-of-speech tagging, and word level alignment.

8. Acknowledgements

I am indebted to Paolo Dongilli for his excellent programming and encoding work.

9. References

Armstrong, Susan, 1996. MULTEXT: Multilingual text tools and corpora. In *Arbeitspapiere zum Workshop Lexikon und Text: Wiederverwendbare Methoden und Ressourcen für die linguistische Erschließung des Deutschen*, Lexicographica. Max Niemeyer Verlag, Tübingen.

Arntz, Reiner and Felix Mayer, 1996. Vergleichende Rechtsterminologie und Sprachdatenverarbeitung — das Beispiel Südtirol. In *Übersetzungswissenschaft im Umbruch*. Gunter Narr Verlag, Tübingen, pages 117–129.

Bowker, Lynne, 1996. Towards a corpus-based approach to terminography. *Terminology*, 3(1):27–52.

Bray, Tim, Jean Paoli, and C.M. Sperberg-McQueen, 1998. Extensible markup language. Technical report, World Wide Web Consortium. Available from http://www.w3.org/TR/REC-xml.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer, 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

Church, Kenneth Ward, 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, Ohio.

Dagan, Ido and Kenneth W. Church, 1997. *Termight*: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12:89–107.

Dagan, Ido, Kenneth W. Church, and William A. Gale, 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*.

Gale, William A. and Kenneth W. Church, 1994. A program for aligning sentences in bilingual corpora. In Susan Armstrong (ed.), *Using Large Corpora*. The MIT Press, pages 75–102. Reprinted from Computational Linguistics, Volume 19, Numbers 1 and 2 (1993).

Goldfarb, Charles F. 1990. *The SGML Handbook*. Oxford University Press.

Habert, B., G. Adda, M. Adda-Decker, P. Boula de Marëuil, S. Ferrari, O. Ferret, G. Illouz, and P. Paroubek, 1998. Towards tokenization evaluation. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada (eds.), *Proceedings of the First International Conference on Language Resources & Evaluation*, volume 1. Granada, Spain.

Heid, Ulrich, Susanne Jauss, and Katja Krüger, 1996. Term extraction with standard tools for corpus exploration — experience from german. In *Proceedings of the 4th International Congress on Terminology and Knowledge Engineering (TKE'96)*. Vienna, Austria: INDEKS Verlag.

Ide, Nancy, 1998. Corpus encoding standard: SGML guidelines for encoding linguistic corpora. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada (eds.), *Proceedings of the First International Conference on Language Resources & Evaluation*, volume 1. Granada, Spain.

Ide, Nancy, Greg Priest-Dorman, and Jean Véronis, 1996. Corpus encoding standard. See http://www.cs.vassar.edu/CES/.

Mayer, Felix (ed.), 1998. *Terminologisches Wörterbuch zur Südtiroler Rechts- und Verwaltungssprache*. Europäische Akademie Bozen.

Melamed, I. Dan, 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL'97)*. Madrid, Spain: Association for Computational Linguistics.

Rubio, Antonio, Natividad Gallardo, Rosa Castro, and Antonio Tejada (eds.), 1998. *Proceedings of the First International Conference on Language Resources & Evaluation*. Granada, Spain: European Language Resources Association.

Simard, Michel, 1998. The BAF: A corpus of English-French bitext. In *Proceedings of the First International Conference on Language Resources & Evaluation*, volume 1. Granada, Spain.

Sperberg-McQueen, C.M. and L. Burnard (eds.), 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago and Oxford: Text Encoding Initiative. Available from http://etext.virginia.edu/tei.html.