

Primary Data Encoding of a Bilingual Corpus

Johann GAMPER Paolo DONGILLI

Abstract

This paper discusses the building of a bilingual corpus of legal and administrative texts, focusing on the encoding of documentation and structural information according to the Corpus Encoding Standard. The corpus is one module in an ongoing research project about (semi-)automatic terminology acquisition at the European Academy Bolzano and will serve as a basis for applying term extraction programs. We will discuss the pieces of information to be annotated as well as lessons learned during this process.

1. Introduction

Due to the equal status of the Italian and the German language in South Tyrol, legal and administrative documents have to be written in both languages. A prerequisite for high quality translations is a consistent and comprehensive bilingual terminology, which also forms the basis for an independent German legal language reflecting the Italian legislation. The first systematic effort in this direction was initiated in 1994 by the Commission for Terminology in cooperation with the scientific area I of the European Academy Bolzano with the goal to compile an Italian/German legal and administrative terminology for South Tyrol (ARNTZ & MAYER 1996; see figure 1).

A few years of experience have shown that manual acquisition of terminological data from texts is a very work-intensive and error-prone task. Recent advances in automatic corpus analysis favored a modern form of terminology acquisition where a corpus is a collection of language material in machine-readable form and computer programs help scanning the corpus for terminologically relevant information, generating lists of term candidates which have to be post-edited by humans. This new form of terminology acquisition will be applied in the CATEx (Computer

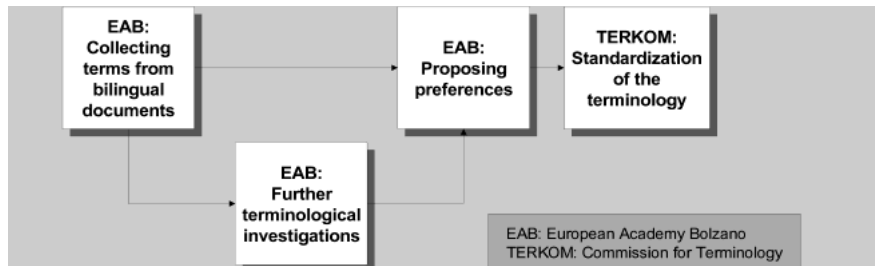


Figure 1: Standardization of the legal and administrative terminology in South Tyrol

Assisted Terminology Extraction) project, which emerged from the need to support and improve, both qualitatively and quantitatively, the manual acquisition of terminological data at the European Academy Bolzano. Thus, the main objective of CATEX is the development of a computational framework for (semi-)automatic terminology acquisition, which consists of four modules:

1. a parallel text corpus;
2. term-extraction programs;
3. a term bank linked to the text corpus;
4. a user-interface for browsing the corpus and the term bank.

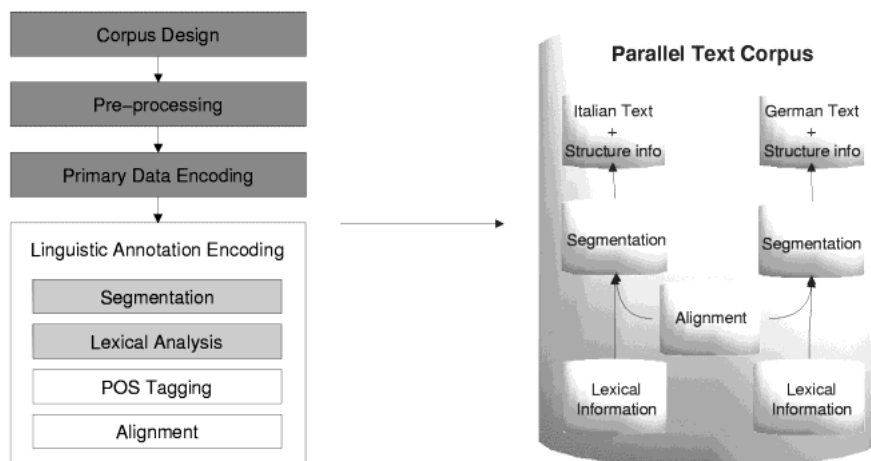


Figure 2: Building scheme of an Italian-German parallel text corpus

Currently, we are building the parallel text corpus (GAMPER 1999). This comprises the following tasks: corpus design, pre-processing, encoding primary data, and encoding linguistic information (see figure 2).

2. Corpus Design and Pre-Processing

The corpus design phase selects a collection of texts which should be included in the corpus. In its current form, our corpus contains only one sort of texts, namely the bilingual version of the most important Italian law codes (see table 1). A particular feature of our corpus, which contains both German and Italian translations, is the structural equivalence of the original text and its translation down to the sentence level. This corpus is one of the largest special language corpora. It contains about 5 million words and 35,898 (66,934) different Italian (German) word forms.

The starting point for the pre-processing phase was the set of law books on paper (see figure 3 for an excerpt of the Civil Code) which was converted into electronic form (figure 4) via optical character recognition (OCR). In the pre-processing phase we corrected (mainly OCR) errors in the raw text material and produced a unified electronic version in order to

LAW BOOKS	WORDS	
	Italian	German
Gesetzbuch des Landes Südtirol	1,791,174	1,623,263
Italienisches Zivilgesetzbuch	247,870	256,071
Nebengesetze zum italienischen Zivilgesetzbuch	113,319	117,061
Italienische Zivilprozeßordnung	110,124	115,962
Italienische Strafprozeßordnung	136,855	143,318
Italienische Notariatsordnung	48,523	51,386
Italienisches Konkursrecht und andere Insolvenzverfahren	32,357	32,311
Einheitstext der Steuern auf das Einkommen	46,193	44,254
Der italienische Verwaltungsprozeß	26,085	26,085
	2,552,500	2,409,711

Table 1: Bilingual corpus figures

simplify the creation of programs for consequent annotations. Among others, we corrected the following errors: superfluous blanks within dates (see figure 3); misread characters such as „I“, „l“ and „1“; missing end-of-paragraph detection for sentences ending exactly at the right margin of the text block. Another step in the pre-processing phase was some kind of standardization of the text. While the Italian original titles are numbered using Roman numbers, most German translations use Arabic numbers. We transformed these Arabic numbers into Roman numbers, e.g. „1. Titel“ has been translated into „Titel I“.

3. Corpus Encoding Overview

Corpus encoding is meant to enrich the raw text material with explicitly encoded information which represents the interpretation of various text elements and features. We decided to apply the Corpus Encoding Standard (CES: IDE et al. 1996) which is an application of SGML (GOLDFARB 1990) and defines a set of guidelines for corpus annotation especially tailored for language engineering. So-called document type definitions (DTDs) are provided, which specify what pieces of information can be encoded in the corpus and what their relative collocation must be. CES distinguishes primary data (raw text material in machine-readable form) and linguistic annotation (information resulting from linguistic analyses of the raw texts).

Our efforts are set out now towards the primary data encoding of the bilingual corpus, as described in the next section. Linguistic annotation will represent our future work.

4. Primary Data Encoding

Primary data encoding comprises the mark-up of documentation and structural information. Documentation information includes global information about the text, e.g. bibliographic information (author, publisher, edition, etc.) and information concerning the distribution of the electronic corpus (institution, address, etc.). Structural annotation covers the mark-up of relevant structural elements in the raw text material. Gross structural mark-up and sub-paragraph mark-up are distinguished. The gross structure

<p>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.</p>	<p>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.</p>
---	---

Figure 3: Civil Code excerpt: Italian original and German translation

of a text consists of elements such as large divisions (chapters, sections, etc.) down to the paragraph level, titles, lists, tables, etc. Sub-paragraph structures include elements like sentences, abbreviations, dates, quotations, references, etc. There are three sequential mark-up levels for primary data encoding which have to be reached in order to follow the CES recommendations.

<p><6>Libro I. <5>Delle persone e della famiglia. <6>Titolo I. <5>Delle persone fisiche. <3>1. (Capacità giuridica). <1>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).<8>1<1><8>1<4> Il comma 3 ' e stato abrogato <8>1<4>n virtù dell'art. 1 R.D.L. 20 gennaio 1944, n. 25 e dell'art. 3 D.Lg.Lt. 14 settembre 1944, n. 287.</p>	<p><6>1. Buch <5>Personen- und Familienrecht <6>1. Titel <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.</p>
---	--

Figure 4: OCR output (<1> normal size, <3> normal size bold, <4> small, <6> large, <5> large boldface, <8> superscript)

Each text is encoded as a `<cesDoc>` element which consists of a header and a body. The header (`<cesHeader>` element) contains the documentation information and the body contains the raw text material and the mark-up for structural information.

The annotation of documentation and structural information serves several purposes. First of all, these pieces of information are necessary to automatically extract the source of terms, e.g. „Codice Civile, art. 320“. Second, structural information is important for the development of a sophisticated user interface to browse the corpus. This is important in our case, since we intend to disseminate the corpus prior to the completion of terminology extraction. A bilingual, sentence aligned corpus provides a valuable resource for translators. Moreover, in a later state the corpus will be linked to the terminological database, hence user-friendly browsing of the corpus becomes important. Finally, documentation information helps to maintain the text corpus.

4.1 Implementation Issues

The general approach we adopted in the pre-processing phase and for structural annotation was to scan the raw texts using a sequence of filters. Each filter adds some small pieces of new information and writes a log file in cases of doubt. The output and the log file are used in turn to improve the filter programs in order to minimize manual post-editing. This modular boot-strapping approach has advantages over huge parameterizable programs: filters are fairly simple and can be partially reused or easily adapted for texts with different formats. The filters have been implemented in Perl, a general purpose interpreted language which, by providing extensive support for regular expression matching, turns out to be a powerful language for such applications.

Most of the gross structural elements can be produced by analyzing the mark-up for text formatting information (angular bracket tags) in the raw text material (OCR output). In this way Level-1 texts are obtained.

Angular brackets from the OCR output files are useful also for the detection of some sub-paragraph level elements such as quotes, lists, items, notes and pointers to notes. The results of this analysis are Level-2

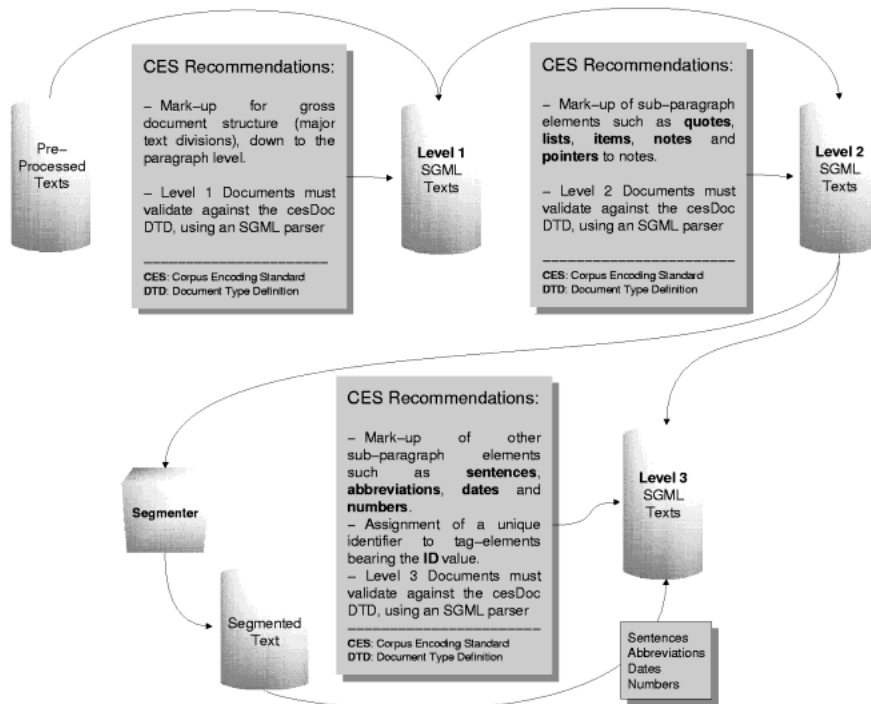


Figure 5: Primary Data Encoding: Level 1, 2 and 3

SGML files. Figure 5 shows the CES recommendations for Level-1 and Level-2 documents.

For the recognition of the remaining sub-paragraph elements in order to reach Level-3 we used the MULTTEXT tokenizer MtSeg (ARMSTRONG 1990; available from <http://www.lpl.univ-aix.fr/projects/multtext>). Tokenization detects structure features (paragraph and sentence boundaries) and particular tokens (abbreviations, dates, digits, etc.), pieces of information that will help us in completing the structural annotation obtaining Level-3 SGML documents. In figure 5 we also show how information collected from the MtSeg output file are merged into a Level-2 SGML file giving a Level-3 text as its result. An excerpt of the Civil Code brought to Level-3 is presented in figure 6.

MtSeg is composed of a series of sub-tools, each devoted to solving a specific problem. The sub-tools perform several processes such as

splitting text at spaces, isolating punctuation, identifying abbreviations, recombining compounds, etc. The rules determining how to treat the different tokens are provided as data to the appropriate sub-tool via a set of language-specific, user-defined resource files and are thus entirely customizable. We had to add new items to the resource files for German and Italian — information we extracted from our texts using ad hoc Perl scripts. The resource files we modified are those which contain abbreviations, compound names and clitics. The segmenter's resource files can be created in a boot-strapping process. By checking the output of the segmenter we verified the entries we already disposed of and saw others that had to be added to the files.

```

<p id=cc2.1.1.1.0.0.0.1.p1>
<s id=cc2.1.1.1.0.0.0.1.p1.s1>Die Rechtsfähigkeit
wird zum Zeitpunkt der Geburt erworben (<num>22</num>
<abbr expan="Verfassung der Republik Italien (Gesetzblatt der Republik
Nr. 298 vom 27.12.1947)">Verf.</abbr>).</s></p>
<p id=cc2.1.1.1.0.0.0.1.p2>
<s id=cc2.1.1.1.0.0.0.1.p2.s1>Die Rechte, die das
Gesetz dem Gezeugten zuerkennt, hängen von der tatsächlichen Geburt ab
(<ref target=cc2.1.1.7.2.1.1.5><num>254</num></ref>,
<ref target=cc2.1.2.1.2.0.0.1><num>462</num></ref>,
<ref target=cc2.1.2.5.3.0.0.3><num>784</num></ref>).
<ptr target=cc2.1.1.1.0.0.0.1.fn1 n=1></s></p>
<note id=cc2.1.1.1.0.0.0.1.fn1 n=1 place=end>
<p id=cc2.1.1.1.0.0.0.1.p3>
<s id=cc2.1.1.1.0.0.0.1.p3.s1>Der dritte Absatz
wurde durch Artikel <num>1</num> des Königlichen Gesetzesdekrets vom
<date ISO8601="1944-01-20">20.1.1944</date>, <abbr expan="Nummer">Nr.</abbr>
<num>25</num>, und durch Artikel <num>3</num> der gesetzvertretenden
Verordnung des Statthalters vom<date ISO8601="1944-09-14">14.9.1944</date>,
<abbr expan="Nummer">Nr.</abbr> <num>287</num>, aufgehoben.</s></p></note>

```

Figure 6: Primary Data Encoding: Level-3 example

After a few boot-strapping sessions on the Civil Code (both German and Italian versions) we were able to tune the resource files at a really satisfactory level. At last we ran MtSeg on a 10% chunk ($\approx 28,000$ words) of the Civil Code and the segmented output was compared with the original code and checked for errors. We found only one error typology for both the German and the Italian codes: 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. „6. Absatz“. The structural equivalence of our parallel texts provides valuable information and allows us to easily detect such segmentation errors.

5. Conclusion

We showed the first results we have achieved in encoding an Italian-German parallel corpus which serves for term extraction purposes; we explained the overall approach and discussed in detail the steps toward a CES-compliant structural annotation of our bilingual documents. Similar projects are underway which also use the CES (e.g. ERJAVEC & IDE 1998). Future work will include the linguistic annotation which enriches the primary data with information resulting from linguistic analyses of these data. We will consider the assignment / disambiguation of lemmas and part-of-speech tags and the alignment of parallel texts, first on the sentence level and later on the word level. We are also working on a sophisticated interface to navigate through parallel documents.

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*, 107-119. Tübingen: Max Niemeyer Verlag.
- ARNTZ, Reiner and MAYER, Felix (1996): Vergleichende Rechtsterminologie und Sprachdatenverarbeitung – das Beispiel Südtirol. In: *Übersetzungswissenschaft im Umbruch*, 117-129. Tübingen: Gunter Narr Verlag.
- ERJAVEC, Tomaž and IDE, Nancy (1998): The MULTEXT-East corpus. In: Antonio RUBIO, Natividad GALLARDO, Rosa CASTRO, and Antonio TEJADA (eds.), *Proceedings of the First International Conference on Language Resources & Evaluation*, Granada / Spain, May 1998, vol. 2, 971-974.
- GAMPER, Johann (1999): Encoding a parallel corpus for automatic terminology extraction. In: *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen / Norway, June 1999, 275-276.
- GOLDFARB, Charles F. (1990): *The SGML Handbook*. Oxford: Oxford University Press.
- IDE et al. (1996): Nancy I., Greg PRIEST-DORMAN and Jean VÉRONIS, Corpus encoding standard. See <http://www.cs.vassar.edu/CES/>.