The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated

Verena Lyding, Elena Chiocchetti
EURAC research
Viale Druso 1, 39100 Bozen/Bolzano - Italy
forename.name@eurac.edu

Gilles Sérasset, Francis Brunet-Manquat
GETA-CLIPS IMAG
BP 53, 38041 Grenoble cedex 9 - France
forename.name@imag.fr

Abstract

Standard techniques used in multilingual terminology management fail to describe legal terminologies as they are bound to different legal systems and terms do not share a common meaning. In the LexALP project, we use a technique defined for general lexical databases to achieve cross language interoperability between languages of the Alpine Convention. In this paper we present the methodology and tools developed for the collection, description and harmonisation of the legal terminology of spatial planning and sustainable development in the four languages of the countries of the Alpine Space.

1 Introduction

The aim of the LexALP project is to harmonise the terminology used by the Alpine Convention, both for internal purposes and for communication among the member states. The Alpine Convention is an international treaty signed by all states of the Alpine territory (France, Monaco, Switzerland Liechtenstein, Austria, Germany, Italy and Slovenia) for the protection of landscape and sustainable development of this mountain area[1]. The member states speak four different languages, namely French, German, Italian, and Slovene and have different legal systems and traditions.

Hence arises the need for a systematization and uniformation of terminology and clear translation equivalence in all four languages. For this reason, the project intends to provide all stakeholders and the wider public with an information system which combines three main components, a terminology data base, a multilingual corpus and the relative bibliographic data base. In this way the manually revised, elaborated and validated (harmonised) quadrilingual information on the legal terminology (i.e. complete terminological entries) will be closely interacting with a facility to dynamically search for additional contexts in a relevant set of legal texts in all languages and for all main legal systems involved.

2 Multilingual legal information system

The information system for the terminology of the Alpine Convention, with a specific focus on spatial planning and sustainable development, will give the possibility to search for relevant terms and their (harmonised or rejected) translations in all 4 official languages of the Alpine Convention in the first module, the term bank. Next to retrieving synonyms and translation equivalents within each legal system, the user will be provided with a representative context and a valid definition of the concept under consideration. Source information will be provided for each text field in the terminological entry.

Via a link from the terminological data base to the second module, the corpus facility, the information system will give the possibility to search the corpus for further contexts.

Finally, both term bank and corpus will be interacting with a third module, the bibliographic database, so as to allow retrieving full information on text excepts cited in the term bank and to store important meta data on corpus documents.

[1] cf. also http: www.alpenkonvention.org

Proceedings of the Workshop on Multilingual Language Resources and Interoperability, pages 25–31,
Sydney, July 2006.

3 Terminological data

3.1 Data categories and motivations

The data categories present in the terminology database allow entering and organising relevant information on the concept under analysis. The term bank interface allows entering of the following terminological data categories: denomination/term, definition, context, note, sources (text fields), grammatical information to the term, harmonisation status, processing status, geographical usage, frequency and domain, according to the appositely elaborated domain classification structure[2] (pull down menus). Again by means of pull-down menus the terminologist will be able to signal to the users which terms are already processed (i.e. checked by legal experts), harmonised or rejected and - most important - to which legal system they belong (the menu geographical usage allows to specify this information). Furthermore it is possible to specify synonyms, short forms, abbreviations etc. in the terminological entry and, if necessary, link them to the relative full information already present in the term bank (however, no direct access to these linked data is possible, this must be done via the search interface). Finally, the terminologist is given the possibility of writing general comments to the entry. At the very end of one language entry the terminologist can decide whether to release the data to the public (by clicking on the button 'finish') or keep it for further fine-tuning (button 'update').

Each term is created in its 'language volume' and described by means of all necessary information. As soon as one or all equivalents in the other languages are available too, the single entries can be linked to each other with the help of an axie (see detailed description below).

Searches can be done for all languages or on a user-defined selection of source and target languages. Presently the database allows global searching in all text fields and filtering by source, author, date of creation, as well as by axie name and ids. Results can be displayed in full form, as a short list of terms only or in XML. Some export/import functions are granted.

As the term bank serves mainly the scope of diffusing harmonised terminology, the four translation equivalents (validated by a group of experts) are displayed together, whereas rejected synonyms are displayed separately for each search language. In this way the user may well look for a non validated synonym and find it in the database but be warned as to which is the preferred term and its harmonised equivalents in the other languages. Figure 1 shows such a situation where the French rejected term "transport intra-alpin" is linked to the harmonised term "trafic intra-alpin".

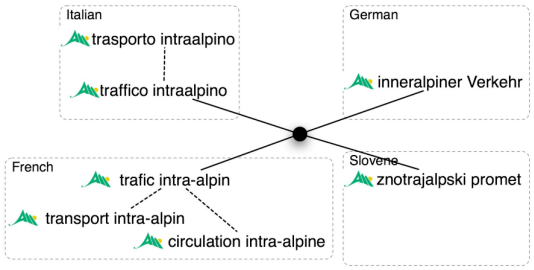


Figure 1: A set of Alpine Convention terms and their relations

3.2 Monolingual data

The LexALP term bank consists in 5 volumes for French, German, Italian, Slovene and English (no data is being entered for this fifth language at the moment), which contain the term descriptions. The set of data categories is represented in an XML structure that follows a common schema.

```
<entry id="fra.trafic_intra-alpin.1010743.e"
       lang="fra"
       legalSystem="AC"
       process_status="FINALISED"
       status="HARMONISED">
  <term>trafic intra-alpin</term>
  <grammar>n.m.</grammar>
  <domain>Transport</domain>
  <usage frequency="common"
         geographical-code="INT"
         technical="false"/>
  <relatedTerm
    isHarmonised="false"
    relationToTerm="Synonym"
    termref="fra.transport_intra-alpin…"/>
  <relatedTerm
    isHarmonised="false"
    relationToTerm="Synonym"
    termref="fra.circulation_intra-…"/>
  <definition>
    [T]rafic constitué de trajets ayant leur
    point de départ et/ou d'arrivée à
    l'intérieur de l'espace alpin.
  </definition>
  <source>Prot. Transp., art. 2 </source>
  <context url="http://www...">
    Des projets routiers à grand débit pour
    le trafic intra-alpin peuvent être
    réalisés, si [...].
  </context>
</entry>
```

Figure 2: XML form of the term 'trafic intra-alpin'

Each entry represents a unique term/meaning. Terms with the same denomination, but belong-

[2] See also 4.1

ing to different legal systems have, de facto, different meanings. Hence, different entries are created. Terms with different denominations but conveying the same 'meaning' (concept) are also represented using different entries[3]. In this case, the entries are linked through a synonymy relation.

Figure 2 shows the XML structure of the French term "trafic intra-alpin", as defined in the Alpine Convention. The term entry is associated to a unique identifier used to establish relations between volume entries.

The example term belongs to the Alpine Convention legal system[4] (code `AC`). The entry also bears the information on its status (harmonised or rejected) and its processing status (to be processed, provisionally processed or finalised).

In addition, a definition (along with its source) and a context may be given. The definition and context should be extracted from a legal text, which must be identified in the source field.

3.3 Achieving language/legal system interoperability

As the project deals with several different legal terms, standard techniques used in multilingual terminology management need to be adapted to the peculiarities of the specialised language of the law. Indeed, terms in different languages are (generally) defined according to different legal systems and these legal systems cannot be changed. Hence, it is not possible to define a common 'meaning' that could be used as a pivot for language interoperability[5]. In this respect, legal terminology is closer to general lexicography than to standard terminology.

In order to achieve language/legal system interoperability we had several options that are used in general lexicography.

Using a set of bilingual dictionaries is not an option here, as we have to deal with at least 16 language/legal system couples (with alpine Convention and EU levels, but without taking into account regional levels). Moreover, such a solution will not reflect the multilingual aspect of the Alpine Convention or the Swiss legal system. Finally, building bilingual volumes between the French and Italian legal systems is far beyond the objectives of the LexALP project.

Another solution would be to use an "Eurowordnet like" approach (Vossen, 1998) where a specific language/legal system is used as a pivot and elements of the other systems are linked by equivalent or near-equivalent links. As such an approach artificially puts a language in the pivot position, it generally leads to an "ethnocentric" view of the other languages. The advantage being that the architecture uses the bilingual competence of lexicographers to achieve multilingualism.

In this project, we chose to use 'interlingual acceptions' (a.k.a. axies) as defined in (Sérasset, 1994) to represent such complex contrastive phenomena as generally described in general lexicography work. In this approach, each 'term meaning' is associated to an interlingual acception (or axie). These axies are used to achieve interoperability as a pivot linking terms of different languages bearing the same meaning.

However, as we are dealing with legal terms (bound to different legal systems), it is generally not possible to find terms in different languages that bear the same meaning. In fact such terms can only be found in the Alpine Convention (which is considered as a legal system expressed in all the considered languages). Hence, we use these terms to achieve interoperability between languages. In this aspect, we are close to Eurowordnet's approach as we use a specific legal system as a pivot, but in our case the pivot itself is generally a quadrilingual set of entries.

These harmonised Alpine Convention terms are linked through an interlingual acception. An axie is a place holder for relations. Each interlingual acception may be linked to several term entries in the languages volumes through `termref` elements and to other interlingual acceptions through `axieref` elements, as illustrated in Figure 3.

```
<axie id="axi..1011424.e">
 <termref
  idref="ita.traffico_intraalpino.1010654.e"
  lang="ita"/>
 <termref
  idref="fra.trafic_intra-alpin.1010743.e"
  lang="fra"/>
 <termref
  idref="deu.inneralpiner_Verkehr.1011065.e"
```

[3] Variants, acronyms, etc. are not considered as different denominations.

[4] Strictly speaking, the Alpine Convention does not constitute a legal system per se.

[5] Consider for instance the difference between the Italian and the Austrian concepts of journalists' professional confidentiality. Whereas the *Redaktionsgeheimnis* explicitly underlines that the journalist can refuse to witness in court in order to keep the professional secret, in Italy the *segreto giornalistico* must obligatorily be lifted on a judge's request. The two concepts have overlapping meanings in the two states, however, they diverge greatly with respect to the behaviour in court.

```
  lang="deu"/>
 <termref
  idref="slo.znotrajalpski_promet.1011132.e"
  lang="slo"/>
 <axieref idref=""/>
 <misc></misc>
</axie>
```

Figure 3: XML form of the interlingual acception illustrated Figure 1

The `termref` relation establishes a direct translation relation between these harmonised equivalents. Then, national legal terms are indirectly linked to Alpine Convention terms through the `axieref` relation as illustrated in Figure 4.

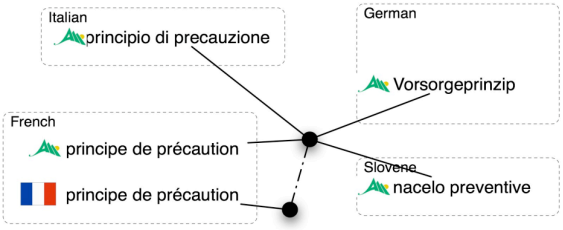


Figure 4: An example French term, linked to a quadrilingual Alpine Convention Term.

4 Corpus

4.1 Corpus content

The corpus comprises around 3000 legal documents of eight legal systems (Germany, Italy, France, Switzerland, Austria, Slovenia, European law and international law with the specific framework of the Alpine Convention,) (see table 1).

AT	CH	DE	FR	IT	SI	AC	EU	INT
612	119	62	613	490	213	38	791	149

Table 1: Corpus documents for each legal system

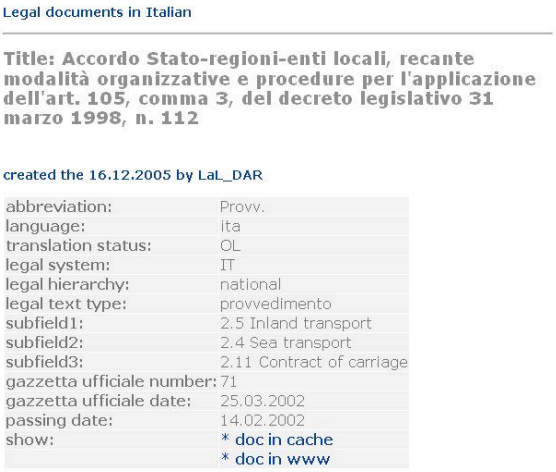
Documents of the supranational level are provided in up to four languages (subject to availability). National legislation is generally added in the national language (monolingual documents) and in case of Switzerland (multilingual documents) in the three official languages of that nation (French, German and Italian).

The documents are selected by legal experts of the respective legal systems following predefined criteria:

- entire documents (no single paragraphs or excerpts etc.);
- strong relevance to the subjects 'spatial planning and sustainable development' as described in art. 9 of the relative Alpine Convention Protocol;
- primary sources of the law for every system at national and international/EU level, i.e. normative texts only (laws, codes etc.);
- latest amendments and versions of all legislation (at time of collection: June – August 2005);
- terminological relevance.

Each document is classified according to the following (bibliographical) categories: full title, short title, abbreviation, legal system, language, legal hierarchy, legal text type, subfield (1, 2 and 3), official date, official number, published in official journal (date, number, page), … The bibliographical information of all documents is stored in a database and can at any time be consulted by the user.

The subfields have been elaborated and selected by a team of legal experts, taking into account the classification specificities followed by the Alpine Convention and the need to classify texts from several different legal systems according to one common structure. For this reason, the legal experts have subdivided the fields spatial planning and sustainable development into 5 main areas, in accordance with the Alpine Convention Protocol dealing with these subjects and subsequently adopted an EU-based model for further subdividing the 5 main topics in such a way that all countries involved could classify their selected documents under a maximum of 3 main items, the first of which must be indicated obligatorily. This classification allows an easy selection of all subsets of documents according to subject field.


Legal documents in Italian

Title: Accordo Stato-regioni-enti locali, recante modalità organizzative e procedure per l'applicazione dell'art. 105, comma 3, del decreto legislativo 31 marzo 1998, n. 112

created the 16.12.2005 by LaL_DAR

abbreviation:	Provv.
language:	ita
translation status:	OL
legal system:	IT
legal hierarchy:	national
legal text type:	provvedimento
subfield1:	2.5 Inland transport
subfield2:	2.4 Sea transport
subfield3:	2.11 Contract of carriage
gazzetta ufficiale number:	71
gazzetta ufficiale date:	25.03.2002
passing date:	14.02.2002
show:	* doc in cache * doc in www

Figure 5: Example of document classification

```
<header
 lang="ita"
 creator="X"
 created="Fri Feb 17 10:45:15 CET 2006">
<h.title>
 Legge_regionale_25974.14_87.txt
</h.title>
<bibID>
 17658
</bibID>
</header>
```

Figure 6: XML-header of corpus documents

```
<text id="17658">
<body id="17658.b">
  <div type="intro" id="17658.b.i">
     <p id="17658.b.i.p1">
       <title id="17658.b.i.p1.ti1">
          LEGGE REGIONALE 15/05/1987, N. 014
          Disciplina dell' esercizio […] di
          fauna selvatica.
       </title>
     </p>
  </div>
  <div type="section" id="17658.b.c0.se1">
     <p id="17658.b.c0.se1.p1">
       <title id="17658.b.c0.se1.p1.ti1">
         Art. 1
       </title>
     </p>
     <p id="17658.b.c0.se1.p2">
       <s id="17658.b.c0.se1.p2.s1">
          1. Sull' intero territorio
          regionale la caccia selettiva
          per qualita', […]
       </s>
       <s id="17658.b.c0.se1.p2.s2">
          a) capriolo: dal 15 maggio al
           15 gennaio;
       </s>
       <s id="17658.b.c0.se1.p2.s3">
          b) cinghiale: dal 15 giugno
          al 15 gennaio;
       </s>
     </p>
     <p id="17658.b.c0.se1.p3">
       <s id="17658.b.c0.se1.p3.s1">
          2. E' ammesso l' uso […]
       </s>
     </p>
  </div>
</body>
</text>
```

Figure 7: XML-structure of corpus document

4.2 Structural organization of corpus data

Collected in raw text format (one file for each legal text) the documents are first transformed into XML-structured files and in a second step inserted into the database.

The XML-annotation is done in compliance with the Corpus Encoding Standard for XML (XCES) [6] . Slightly simplified, the provided schema[7] serves to add structural information to the documents. Each text is segmented into subsections like: preamble, chapter, section, paragraph, title and sentence. Furthermore, a link to the classification data (bibliographic data base) is inserted and, in case of multilingual documents, alignment is done at sentence level.

The XML-annotated documents hold all the information needed for the insertion into the corpus database, such as structural mark-up and bibliographical information. The full text documents are transformed into sets of database entries, which can be imported into the database.

4.3 Technical organization of corpus data

Following the *bistro* approach as realized for the Corpus Ladin dl'Eurac (CLE) (Streiter et al. 2004) the corpus data is stored in a relational database (PostgreSQL). The information present in the XML-annotated documents is distributed among four main tables: document_info, corpus_words, corpus_structure, corpus_alignment.

The four tables can be described as follows:

document_info: This table holds the meta-information about the documents; each category (like full title, short title, abbreviation, legal system, language, etc.) is represented by a separate column. For each legal document one entry (one row) with unique identification number is added to the table. These identification numbers are cited in the XML-header of the corpus documents.

corpus_words: This table holds the actual text of the collected documents. Instead of storing entire paragraphs as it was done during the creation of CLE, for this corpus a different approach is being tested. Every annotated text is split into an indexed sequence of words, starting with counter one. Once inserted into the database a text is stored as a set of tuples composed of word, position in text and document id (as a reference to the document information).

corpus_structure: This table holds all information about the internal structure of the documents. Titles, sentences, paragraphs etc. are stored by indicating starting and ending point of the section. For each segment a tuple of segment type, segment id, starting point (indicated by the index of the first word), ending point (indicated by the index of the last word) and document id is added.

corpus_alignment: This table defines the alignment of multilingual documents. By providing one column for each language the texts are aligned via the document ids or via the ids of single segments.

[6] http://www.cs.vassar.edu/XCES/

[7] http://www.cs.vassar.edu/XCES/schema/xcesDoc.xsd

The tables are interconnected by explicitly stated references. That means that the columns of one table refer to the values of a certain column of another table. As shown in figure 8 all tables hold a column *document_id* that refers to the document id of the table document_info. Furthermore, the table corpus_structure holds references to the column *position* of the table corpus_words.

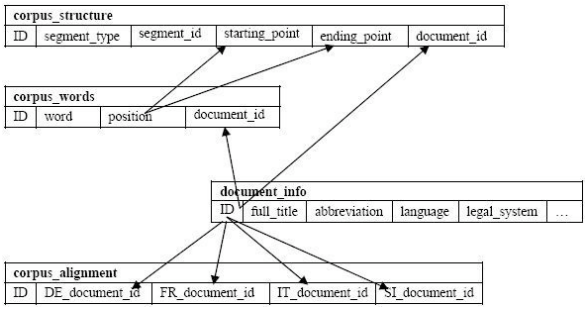


Figure 8: Interconnection of tables

5 Searching the corpus

Due to the fine-grained classification (see section 4.1) and the structural mark-up (see section 4.2) of all corpus documents, corpus searches can be restricted in the following ways:

- by specifying a subset of corpus documents over which the search should be carried out (e.g. all documents of legal system CH with language French);
- by choosing the type of unit to be displayed (whole paragraphs <p>, sentences <s>, titles <title>, …);
- by searching for whole words only (exact match) or parts of words (fuzzy match);
- by restricting the number of hits to be displayed at a time.

For searches in multilingual documents it will be possible to search for aligned segments, specifying search word as well as target translation. For example, the user could search for all alignments of German-Italian sentences that contain the word *Umweltschutz* translated as *tutela ambientale* (and not with *protezione dell'ambiente*).
Figure 9 shows a simple interface for searching monolingual documents.

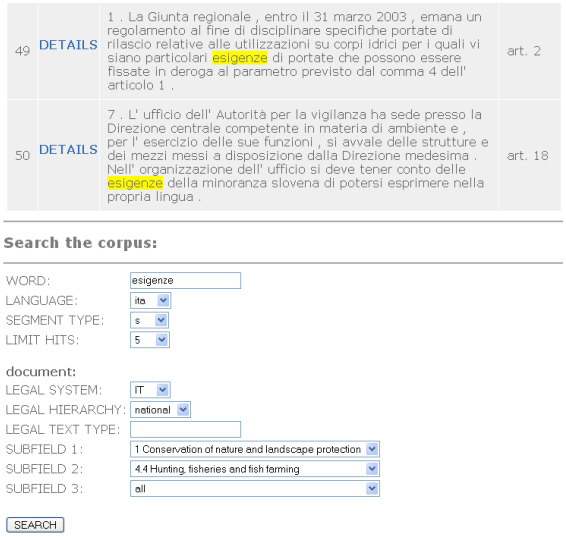


Figure 9: Example search over monolingual documents

6 Interaction term bank and corpus

Term bank and corpus are independent components which together form the LexALP Information System.

The interaction between corpus and term bank will concern in particular 1) corpus segments used as contexts and definitions in the terminological entries, 2) short source references in the term bank (and the associated sets of bibliographical information) and 3) legal terms.

6.1 Entering data into term bank

When adding citations to a term bank entry, the relative bibliographic information will automatically be counterchecked with the contents of the bibliographical database. In case the information about the cited document is already present in the DB, a link to the term bank can be added. Otherwise the terminologist is asked to provide all information about the new source to the bibliographic database and later create the link.

Next to static contexts and definitions present for each terminological entry, each entry will show a button for the dynamic creation of contexts. Hitting the button will start a context search in the corpus and return all sentences containing the term under consideration.

6.2 Searching the corpus

When searching the corpus the user will have the opportunity to highlight terms present in the term bank. In the same way standardised or rejected terms can be brought out. Via a link it will then

be possible to directly access the term bank entry for the term found in the corpus.

In general each corpus segment is linked to the full set of bibliographic information of the document that the segment is part of. Accessing the source information will lead the user to a detailed overview as shown in figure 4.

7 Conclusion

In this paper, we have presented the LexALP information system, used to collect, describe and harmonise the terminology used by the Alpine Convention and to link it with national legal terminology of the alpine Convention's member states. Even if we currently give a specific focus on spatial planning and sustainable development, the project is not restricted to these fields and the methodology and tools developed can be adapted to legal terminology of other fields.

In this paper we also proposed a solution to the encoding of multilingual legal terminologies in a context where standard techniques used in multilingual terminology management usually fail.

The terminology developed and the corpus used for its development will be accessible online for the stakeholders and the wider public through the LexALP information system.

Acknowledgements

The LexALP research project started in January 2005 thanks to the funds granted by the INTERREG IIIB "Alpine Space" Programme, a Community Initiative Programme funded by the European Regional Development Fund.

References

Gilles Sérasset. 1994. *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA*. In Makoto Nagao, editor, COLING-94, volume 1, pages 278—282, August.

Streiter, O., Stuflesser, M. & Ties, I. (2004). CLE, an aligned Tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface, *LREC 2004, Workshop on "First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation"* Lisbon, May 24, 2004.

Vossen, Piek. 1998. Introduction to EuroWordNet. In Nancy Ide, Daniel Greenstein, and Piek Vossen, editors, Special Issue on EuroWordNet, *Computers and the Humanities*, 32(2-3): 73-89.

Wright, Sue Ellen 2001. *Data Categories for Terminology Management.* In Sue Ellen Wright & Gerhard Budin, editors, Handbook of Terminology Management, volume 2, pages 552-569.