Interoperability in Action for Multilingual Legal Terminology: The LexALP Project

Gilles Sérasset (gilles.serasset@imag.fr) and Francis Brunet-Manquat (francis.brunet-manquat@imag.fr)  
GETA-CLIPS-IMAG – CNRS & University Grenoble 1  
BP 53 – 38051 Grenoble Cedex 9 – France

Elena Chiocchetti (elena.chiocchetti@eurac.edu) and Verena Lyding (verena.lyding@eurac.edu)  
EURAC Research – Viale Druso 1 – 39100 Bozen/Bolzano – Italy

Abstract. The LexALP project’s goal is to harmonise the terminology on spatial planning and sustainable development used within the Alpine Convention, so that the member states are able to cooperate and communicate efficiently in the four official languages (French, German, Italian and Slovene). To this purpose, LexALP uses the Jibiki platform to build a term bank for the contrastive analysis of the specialised terminology used in six different national legal systems and four different languages. In this paper we show how we achieve interoperability between languages/legal system, between different dictionaries and between corpus and term banks.

Keywords: Multilingual Legal Terminology, Dictionary Writing System

1. Introduction

One of the most time-consuming hindrances to supranational law drafting and convention negotiation is the lack of understanding among negotiators and technical writers. This is not only due to the fact that different languages are involved, but mainly to the inherent differences in the legal systems. Countries that speak the same language (like France and part of Switzerland) may use the same word to represent different legal concepts, as defined in their respective legal traditions. The same concept may be referred to in different ways according to the

1 Legal Language Harmonisation System for Environment and Spatial Planning within the Multilingual Alps
2 http://www.convenzionedellealpi.org
1 E.g.: In the German-speaking province of Bolzano Italy the Landeshauptmann is the president of the provincial council, with much more limited competence that the Austrian Landeshauptmann, who is head of one of the states (Bundesland) that are part of the Austrian federation.
legal system\textsuperscript{2}. Also, terms that may superficially seem to be translations of each other can represent different legal notions\textsuperscript{3}.

In order to concretely address these problems, several institutions representing translators, terminologists, legal experts and computational linguists joined in the LexALP project, co-funded by EU’s INTERREG IIIb Alpine Space programme. The objective of the project is to compare the specialised terminology of six different national legal systems (Austria, France, Germany, Italy, Switzerland and Slovenia) and three supranational systems (EU law, international law and the particular framework of the Alpine Convention) in the four official languages of the Alpine Convention, which is an international framework agreement signed by all countries of the Alpine arc and the EU. This contrastive analysis serves as a basis for the work of a group of experts (the Harmonising Group) who will determine translation equivalents in French, Italian, German and Slovene (one-to-one correspondence) in the fields of spatial planning and sustainable development for use within the Convention, thus optimising the understanding between the Alpine states at supranational level.

The tools that are to be developed for these objectives comprise a corpus bank and a term bank. The corpus bank is developed by adapting the bistro system (Streiter et al., 2006; Streiter et al., 2004). The term bank is based on the Jibiki platform (Mangeot et al., 2003; Séraasset, 2004).

This project prototypically shows that interoperability is an absolute requirement for the development of a fine grain multilingual lexical/terminological database. In this paper we will focus on three axis of interoperability. The first section presents how we achieve interoperation between languages and legal systems by the use of a pivot macro-structure inspired by the Papillon multilingual lexical database. Then, the second section focuses on the interoperation between a generic dictionary writing system and a particular terminological structure and, through this, interoperation between different dictionary structures. Finally, we will present the interoperation between corpus bank and term bank which is crucial for the definition of terms and for their harmonisation.

\textsuperscript{2} See for instance the European Union use of \textit{chien drogue} while French legislation calls them \textit{chien renifleur}.

\textsuperscript{3} For example, in Italy an \textit{elezione suppletiva} is commonly held whenever an elected deputy or senator either resigns or dies. In Germany in such cases the first non-elected candidate is called to parliament. \textit{Ersatzwahlen} are a rare phenomenon, foreseen in some very specific cases.
2. The LexALP Terminology Structure

2.1. Overview

The objective of the LexALP project is to compare the specialised terminology of six different national legal systems and three supranational systems in four different languages, and to harmonise it, thus optimising communication between the Alpine states at supranational level. To achieve this objective, the terminology of the Alpine Convention is described and compared to the equivalent terms used in national legislation. The resulting terminology entries feed a specific term bank that will support the harmonisation work.

As the project deals with legal terms, which refer to concepts that are proper of the considered national law or international convention, equivalence problems are the norm, given that concepts are not “stable” between the different national legislations. Standard terminology techniques for other fields can not be applied to the field of law, where the standardisation approach (Felber, 1987; Felber, 1994) is not applicable. For this, we chose to use “interlingual acceptions” (aka. axies) as they are defined in the Papillon dictionary (Sérasset, 1994) to represent the equivalence links between concepts of the different legal systems (Arntz, 1993).

![Figure 1. An Alpine Convention concept in four languages.](image)

The example given in figure 1\(^4\) shows a concept defined in the Alpine Convention. This concept has the same definition in the four languages of the Alpine Convention but is expressed by different denominations. The Alpine Convention uses the terms “mesures de compensation” or “mesures compensatoires” which are identified as synonyms by the terminologist.

This illustrates the first goal of the LexALP project. In different texts, the same concept may be realised by different terms in the same

\(^4\) In these figures, each term is represented by its denotation, along with its language (fra, deu, ...) and its legal system (AC for Alpine Convention, EU for European texts, INT for other international laws, FR for France, ...
language. This may lead to inefficient communication. Hence, a single term has to be determined as part of a *harmonised quadruplet of translation equivalents*. The other denominations will be represented in the term bank as non-harmonised synonyms in order to direct drafting and translating within the Alpine Convention towards a more clear and consistent terminology use for interlingual and supranational communication.

When terms bound to other legal systems are considered as representing the same concept, they are linked to the same central node. Figure 2 illustrates a case where terms in other national and supranational legal systems denotes the same concept. Each of these terms are described independently (as they do not share their definitions or context).

*Figure 2. A multilingual/multilegal set of terms sharing the same meaning.*

When terms of other languages/legal system do not share the same meaning, they are linked to a central node of their own. These nodes may be linked in order to represent “translatability” when the meaning is not preserved. Figure 3 illustrate such a case, where German terms were considered as different, but acceptable (or observed) as translations of the Alpine Convention term in particular circumstances.

This illustrates the second goal of the project, which is to help with the fine comprehension of the Alpine Convention and with the detailed knowledge necessary to evaluate the implementation and implementability of the convention in the different legal systems.

As a by-product of the project, one can see that there is an indirect relation between concepts from different national legal systems (by way of their respective relation to the concepts of the Alpine Convention). However, establishing these indirect relations is not one of the main objectives of the LexALP project and would require more direct contrastive analysis.
2.2. MACRO- AND MICRO- STRUCTURES

The LexALP term bank consists in 5 volumes (for French, German, Italian, Slovene and English) containing all term descriptions (grammatical information, definition, contexts etc.). The translation links are established through a central acception volume. Figure 3 show examples of terms extracted from the Alpine Convention, as well as inter-lingual relations by way of acceptions.

2.2.1. Lexies

All language volumes share the same microstructure, represented in XML, and illustrated figure 4.

The term entry is associated to a unique identifier used to establish relations between volume entries. Each term entry belongs to one (and only one) legal system. The example term belongs to the Alpine Convention legal system\(^5\) (code AC). The set of known legal systems includes of course countries belonging to the Alpine Space (Austria, France, Germany, Italy, Slovenia and Switzerland\(^6\)) but also international treaties or conventions. The entry also bears the information

---
\(^5\) Strictly speaking, the Alpine Convention does not constitute a legal system per se.
\(^6\) Also Liechtenstein and Monaco are parties to the Alpine Convention, however, their legal systems are not terminologically processed within LexALP.
on its status (harmonised or rejected) and its process status (to be processed, provisionally processed or finalised).

The term itself and its part of speech is also given, with the general domain to which the term belongs, along with some usage notes. In these usage notes, the attribute geographical-code allows for discrimination between terms defined in national (or federal) laws and terms defined in regional laws as in some of the countries involved legislative power is distributed at different levels.

Then the term may be related to other terms. These relations may lead to simple strings of texts or to autonomous term entries in the dictionary by the use of the termref attribute. The relation itself is specified in the relationToTerm attribute. The current schema allows for the representation of relations between concepts (synonymy, hyponymy and hyperonymy), as well as relations between graphies (variant, abbreviation, acronym, etc.).

Then, a definition and a context may be given. Both should be extracted from legal texts, which must be identified in the source field.

2.2.2. Aaxies
An interlingual acception (or axie) is a place holder for relations. Each interlingual acception may be linked to several term entries in the language volumes through termref elements and to other interlingual acceptions through axieref elements.
2.3. INTEROPERABILITY BETWEEN LANGUAGES/LEGAL SYSTEMS

Though it remains a terminology project, the LexALP system deals with complex language/legal system interoperation as (unlike common terminology project) there is no common set of concept to be used as a connection point between languages. Here, we used interlingual acceptions (axies) inspired by multilingual lexicography projects.

Directly inspired by the Papillon project, these axies have many in common with “concepts” of the EDR (Japan Electronic Dictionary Research) project (EDR, 2002; Miyoshi et al., 1996). They are used as a pivot and represent the union of concepts of all languages/legal systems involved in the project. Unlike EuroWordNet synsets (Vossen, 1998), they are not bound to a particular language. Hence, we provide a finer interoperability between language as we do not introduce artificial contrastive problems. As an example, figure 3 shows that the Slovene lexies zavarovana živalska vrsta (slv, SL) and živali zavarovanih vrst (slv, SL) do indeed represent the same Slovene concept, even if it is not a concept of our central legal system.

3. The LexALP term bank

Building such a term bank can only be envisaged as a collaborative work involving terminologists, translators and legal experts from all the involved countries. Hence, the LexALP consortium has set up a centralised information system that is used to gather all textual and terminological data.

This information system is organized in two main parts. The first one is dedicated to corpus management. It allows the users to upload legal texts that will serve to bootstrap the terminology work (by way of candidate term extraction) and to let terminologists find occurrences of the term they are working on, in order for them to provide definitions or contexts (see section 4).

The second part is dedicated to terminology work per se. In this section, we show how interoperation between a generic dictionary writing system (jibiki) and a specific dictionary structure is achieved in order to efficiently create the term bank of the LexALP information System

3.1. JIBIKI PLATFORM OVERVIEW

The Jibiki platform has been designed to support the collaborative development of multilingual dictionaries. This platform is used as the basis of the Papillon project web site \(^7\).

\(^7\) http://www.papillon-dictionary.org/
This platform offers several services to its users:

- access to many different dictionaries from a simple or an advance search form,
- creation and editing of dictionary entries.

What makes the Jibiki platform quite unique is the fact that it provides these services regardless of the dictionary structure. In other words it may be used by any dictionary builder to give access and collaboratively edit any dictionary, provided that the resulting dictionary will be freely accessible online.

For concision reasons, we will only address the dictionary access service.

### 3.2. Dictionary access

The first main service consists in browsing the currently developed dictionary. It consists in two different query interfaces (simple and advance search) and a unique result presentation interface. The specified search may only be performed if the system:

- knows in which volume the search is to be performed,
- knows where, in the volume entry, the headword is to be found,
- is able to produce a presentation for the retrieved XML structures.

As the Jibiki platform is entirely independent of the underlying dictionary structure (which makes it highly adaptable), additional metadata is required. These pieces of information are to be found in the mandatory dictionary descriptor. It consists in a structure defined in the Dictionary Metadata Language (DML), as set of metadata structures and a specific XML namespace defined in (Mangeot, 2001).

Figure 5 gives an excerpt of this descriptor. The metadata first identify the dictionary by giving it a name and a type. In this example the dictionary is a pivot dictionary (DML also defines monolingual and bilingual dictionary types). The descriptor also defines the set of source and target languages. Finally, the dictionary is defined as a set of volumes, each volume being described in another file. As the LexALP dictionary is a pivot dictionary, there should be a volume for the artificial language axi, which is the pivot volume.

Figure 5 also shows an excerpt of the description of the French volume of the LexALP dictionary. After specifying the name of the dictionary, the descriptor provides a set of cdm-elements. These elements
The LexALP project

Figure 5. Excerpts of the dictionary and volume descriptors

are used to identify standard dictionary elements (that can be found in several dictionaries) in the specific dictionary structure. For instance, the descriptor tells the system that the headword of the dictionary (cdm-headword) is to be found by applying the specified xpath\(^8\) to the dictionary structure.

With this set of metadata, the system knows that:

- requests on French should be directed to the LexALP\(_{\text{fra}}\) volume,
- the requested headword will be found in the text of the term element of the volume entry element,

Hence, the system can easily perform a request and retrieve the desired XML entries. The only remaining step is to produce a presentation for the user, based on the retrieved entries. This is achieved by way of a xsl\(^9\) stylesheet. This stylesheet is specified either on the dictionary level (for common presentations) or on the volume level (for volume specific presentation).

\(^8\) an xpath is a standard way to extract a sub-part of any XML structure

\(^9\) XSL is a standard way to transform an XML structure into another structure (XML or not).
This mechanism allows for the definition of presentation outputs in **xhtml** (for online browsing) or for presentation output in **pdf** (for dictionary export and print).

### 3.3. Interoperability between dictionaries

As the Jibiki platform is independent of the dictionary structure, it may be used to access, simultaneously, several different dictionaries, even if they do not share a common structure.

Using a simple common dictionary structure (the CDM) used to identify elements that are usually found in dictionaries, it allows for the parallel browsing of several dictionaries. Hence it offers a first, simple, answer to interoperability of differently structured dictionaries. This answer is quite limited though, as it allows for alignment of dictionary structures and sub-structure, but it does not cope with differences in granularity of the dictionaries.

### 4. The LexALP corpus bank

#### 4.1. Corpus content and organisation

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 I).

<table>
<thead>
<tr>
<th></th>
<th>AT</th>
<th>CH</th>
<th>DE</th>
<th>FR</th>
<th>IT</th>
<th>SI</th>
<th>AC</th>
<th>EU</th>
<th>INT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>612</td>
<td>119</td>
<td>62</td>
<td>613</td>
<td>490</td>
<td>213</td>
<td>38</td>
<td>791</td>
<td>149</td>
</tr>
</tbody>
</table>

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).

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.
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)\(^\text{10}\). Slightly simplified, the provided schema\(^\text{11}\) serves to add structural information to the documents. Each text is segmented into sub-sections 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.

4.2. INTEROPERABILITY BETWEEN CORPUS AND TERM BANKS

This axis of interoperability is, by far, the less developed of the system. Such an interoperability is crucial. Corpus data is used for term extraction. Then term description makes heavy use of the bibliographical data (to identify the sources of definition or examples). Also the LexALP information system provides ways to go from terms to corpus and back. On this aspect, interoperability is achieved by ad-hoc tools and can not (yet) be achieved by generic matters. The next development of the project will focus on this aspect.

5. Conclusion

In this article we give some details on the way the LexALP project achieve interoperability on three different axis.

First, we describe the lexical structure that allows for interoperability between languages/legal systems. This pivot structure is based on “axies”. Unlike the EuroWordNet approach, no natural language is arbitrarily considered as central, leading to a more accurate description of contrastive phenomena.

Then, we show how we handle interoperability between different dictionaries by way of a simple standard structure on which each dictionary is projected\(^\text{12}\). Such a projection allows for the use of a generic online dictionary writing system, Jibiki which has already been used for the Papillon multilingual dictionary\(^\text{13}\), or the GDEF (Grand Dictionnaire Estonien-Français) project\(^\text{14}\) an Estonian French bilingual dictionary.

\(^{10}\) \url{http://www.cs.vassar.edu/XCES/}

\(^{11}\) \url{http://www.cs.vassar.edu/XCES/schema/xcesDoc.xsd}

\(^{12}\) For the Jibiki platform, we use the Common Dictionary Mark-up format, but the very same technique may be based on other standard dictionary structures

\(^{13}\) \url{http://www.papillon-dictionary.org/}

\(^{14}\) \url{http://estfra.ee/}
Finally, we show that more work is to be done on interoperability between corpus and term banks.

References


Streiter, O., L. Voltmer, I. Ties, and N. Ralli: 2004, ‘BISTRO, the online platform for terminology management: structuring terminology without entry structures’. In: The translation of domain specific languages and multilingual terminology, No. 3 in Linguistica Antverpiensia New Series. Hoger Instituut voor Vertalers en Tolken, Hogeschool Antwerpen.
