Multilingual Processing in Eastern and Southern EU Languages

Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation

Edited by

Cristina Vertan and Walther v.Hahn

CAMBRIDGE SCHOLARS PUBLISHING

Multilingual Processing in Eastern and Southern EU Languages:
Low-Resourced Technologies and Translation,
Edited by Cristina Vertan and Walther v.Hahn

This book first published 2012

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-3878-0, ISBN (13): 978-1-4438-3878-8

This volume is supported by

META-NET, a European Network of Excellence

TABLE OF CONTENTS

Chapter Three: Multilingual Applications

INTRODUCTION: TRANSLATION DIFFICULTIES AND INFORMATION PROCESSING PROBLEMS WITH EASTERN EUROPEAN LANGUAGES

CRISTINA VERTAN AND WALTHER V.HAHN

It is still popular today to blame machine translation (MT) for poor translations of literary texts. However, even inelegant translations are an industrial factor in producing MT software, in selling multilingual retrieval for relevance scanning or in opening markets by issuing simple foreign-language descriptions. Information retrieval (IR) technologies are effective even if their degree of linguistic correctness is low.

The success story of Machine Translation is partly owed to some simplifications, which made its start-up easier (leaving aside the political presetting of English-Russian translations). Simplifying reality was a promising approach, because the reduction of parameters from syntax, morphology, and domain coverage formed the basis for the demonstration of MT's feasibility.

Moreover, the statistical approach in MT nourished the hope that reasonable results for English can be seen as evidence for the fact that MT can be done with similar quality for any other language.

In subsequent decades experiments were performed with numerous other language pairs around the world including languages even, for which detailed linguistic knowledge was unavailable . The goal of these scientific and industrial research efforts was mainly to estimate the quality and costs of acceptable MT products for the commercially meaningful language pairs. The same holds true for multilingual information retrieval and multilingual information processing on other fields.

With the ever growing number of language pairs for which customers require cost-efficient processing, four aspects became clear:

1. There are domains and language pairs for which not even human translation/IR is available, e.g., financial law texts from Finish to, say, Hausa. The question remains how to obtain these at all.

2. There is no representative bilingual data collection (a "corpus") for these language pairs at all. Statistical approaches hence will not be feasible within the next 5-10 years. How to obtain inexpensive translations in the mid-term for these "low resourced" languages?
3. Many languages (e.g., Hausa) have more than one writing system or changing orthography (compare the post-reform rules for German orthogography that are in force since 2006). This poses the challenge of how to obtain homogeneous corpora.
4. Multinational or global companies need language processing for promotion, local instruction, or contracts, that affords legally binding results.

The optimism of the pioneering years has yielded to scepticism regarding general recipes for multilingual processing such as translation, even for the traditional Western languages. In Europe, the expansion of the EU additionally demonstrated that democratic co-operation requires a huge work load of translation and bilingual information processing among today's 23 official languages. The sheer number of languages, their diverse linguistic structure and their different public use are reasons enough to give up some of the starting assumptions and simplifications of the first decades.

This publication discusses and demonstrates the rather different situations in Europe with regards to cross-lingual processing tasks in an English and American context along the following dimensions:

1. Languages

There are 230 spoken languages in Europe. Most of them have a long common history in the Indo-European paradigm. Even among the 23 official languages of the EU there are the Finno-Ugric official languages Finnish, Estonian and Hungarian. The Turkic and Mongolic families also have several European members, while the north Caucasian and Kartvelian families are important in the south-eastern border of geographical Europe. The Basque language of the Western Pyrenees is an isolated language, unrelated to any other language group in Europe. Much less known even to Central European citizens is the existence of a European Semitic language, the Maltese, written in Latin letters.

In the current volume we decided to refer only to the official EU languages, as representatives of most of the families enumerated above. Additionally, due to European integration there is an increased need in translation and cross-lingual management of documents in these

languages. We hope that the some solutions presented here can be applicable also for non-official and minority languages of the EU.

Even the simple enumeration of the language families encountered in Europe already reveals the existence of major graphemic, phonetic and structural differences amongst them. The aim of this volume is not to investigate these differences from a linguistic point of view, but rather to insist on those discrepancies that trigger challenges for any translation system or cross-lingual/multilingual application. In this sense the following aspects are of relevance:

1.1 Writing differences

Although Europe has no unusual iconographic or syllabic writing systems but only phonographic paradigms, there are nevertheless problems with gathering homogenous bilingual language resources, i.e., training material for statistical approaches.

Cyrillic transliterations for named entities (NEs) follow four different (target language independent) transliteration schemata and numerous (target language dependent) transcriptions. As an example, consider the (operating system dependent) specific encodings for Bulgarian Cyrillics in contrast to Russian encoding. A similar situation exists for Arabic NEs in Maltese. The transliteration is not always standardised, which often leads to data sparseness. One word transliterated in three different ways will be identified in fact as three different words. Moreover, there is a problem with older electronic resources that were developed before the introduction of the Unicode character set: Many languages adopted transliteration simplifications that induce undesired ambiguities. For example the Romanian word for “goose” contains two diacritics “gâscă”. A corpus collected before the introduction of the Unicode system would simplify this word to “gasca”, which may be read also as “gaşcă” denoting a group of young people.

1.2 Variety of Linguistic Structure

European languages differ centrally in their use of pronouns and articles. So called "pro-drop" languages like Italian do not express the 1st person sg. pronouns explicitly, but mark them morphologically, whereas non-pro-drop languages like German have to add the pronoun explicitly in examples like "Ich gehe zu ihm" (I am going to him). Even more difficult in this sense is Hungarian, where lots of particles are only attached as

morphemes (összerakhatatlanságukért = "for their quality of not being easy to put together")[1].

Grammatical gender is present or not (English, Basque), is expressed in noun endings (Italian, mostly), or not (German, mostly), affects other words by agreement (Spanish, French) affects the demonstratives (Italian), or not (Greek), is additionally marked by articles (German, but not unambiguously), or not (Bulgarian, the Baltic languages). Articles are in use for the three-gender system (German), or two genders (Maltese), or the middle (Romanian, common singular for masc. and ntr.) attached to the end of a word (Romanian), or separated in front of the noun (French). Moreover, grammatical and natural gender have an unclear relation in most languages.

All these are major challenges especially in machine translation (MT) whenever the target language is more productive in pronouns or articles than the source. Rule based MT needs a deep linguistic analysis module and often the involvement of large knowledge bases in order to infer the correct target pronoun, while corpus-based MT cannot cope with this problem at all. In most cases the translation lacks not only the correct pronoun but also all derived information such as the correct inflection of the dependent nouns, adjectives and verbs.

To express definiteness, some languages use articles, while others express it by word order, which normally gets lost in surface-form statistical MT systems.

The word order of adjective and noun is semantically relevant in Spanish, restricted in German and fixed in English. Also the position of a verb in the sentence varies among language families. This is a real challenge not only for translation systems but also for multilingual tools that try to apply the same analysis technique to several structurally different languages. Rule-based tools lack a substantial number of common rules. Statistical methods, on the ther hand, require the availabilty of huge non-sparse data covering all these phenomena.

Word composition plays a major role in many EU languages and the order of components is significant. Sometimes logical particles must be inserted for correct translation. Distant verb particles in German are very difficult to differentiate from prepositions when only statistical methods are being applied. Again, such particularities constitute challenges not only in translation but for any preprocessing step in cross-linguistic processing.

[1] This example is owed to Merényi Csaba from MorphoLogic

1.3 Contact

All these languages have been in extensive contact with each other over time with the result, that additional irregularities were introduced. In Romanian, for example, one third of the vocabulary stems from Russian, Hungarian and German and has only been assimilated superficially. This means that the graphemic rules for Romanian are not homogenous. The contact, however, differs from language to language. Compare Romanian-Italian and Slovene-Italian contacts, e.g. whichdiffer with respect to the historical time, when the contact was established. Italy influenced Slovenia much more through Venetian than through modern Italian.

Borrowings were often done only partially so not all semantics is preserved. A special situation is encountered on the Balkan Peninsula where vocabulary related to food, e.g., old weapons or customs are usually shared by neighbouring communities but not necessarily by whole countries. For example, a lot of regional words in the North-Western part of Romania (Transylvania) are in common with Hungarian and German, but unknown in the Southern part of the country, which otherwise uses more words shared with Turkish and Bulgarian.

This constitutes a real challenge for modern retrieval systems which make use of ontologies. Building a language-independent ontology is an extremely difficult task, and even word-based semantic networks are highly problematic. A series of papers published over the last years report the difficulties in adapting the English Word-Net to each of the Balkan languages, and the challenge of homogenisation amongst these Word-Nets.

These considerations, however, touch upon cultural differences, that are addressed in the following paragraphs of this paper. We demonstrate with a few observations that behind the language differences in the EU there are many more cultural differences than between two regions of one country.

2. Text Structures, Forms and Formats

Two textual peculiarities of European publications are very confusing in corpora:

European texts often quote passages in a foreign language such as English or other European languages, because of a close contact with that language. Translated quotations from web pages, hence, are not always in the correct language, as an MT tool has been used.

A typical text form for an application, for an objection, for an expertise etc. differs not only lexically but also in style and form between European countries.

3. Cultures of Textuality

In Europe the influence of English varies from country to country. A comparison of German and French shows that an official inhibitive language policy influenced the borrowings to a high degree in France. Looking to Hungarian one can observe that the French policy more or less succeeded in most technical fields, whereas in Hungary two different medical nomenclatures exist that in practice and international information exchange conflict severely.

Irrespectively of a country's official language policy, the language policy of companies is also changing dramatically. A recent study observed that in Germany slightly more than 50% of all companies use German as the only business language, another 20% use German and English or only English, respectively. Less than 4% use other languages.

In general, technical fields in countries have a different degree of textualisation dependent on technologies, which have a higher or lower distance to text production and use, e.g. carpet weaving, vs. violin making vs. ecological food supplier vs. photo-copying or services.

Correspondingly, the reference to computerised texts, e.g. interactive web forms or download resources of public service differs very much between European countries, say, between Finland and Poland. While web forms must be explained even for language minorities, paper forms are issued by offices, where misunderstandings may be resolved in direct contact.

4. Commercial Market Value

The possible market value of multilingual or cross-lingual technologies clearly depends on the mere number of publications accessible and resulting from trade and industry. If you compare a Slavonic language minority like the Sorbs in Germany to Polish speakers in Poland, it becomes clear that the industrial expansion in Poland and exports abroad result in a disproportionately more extensive language and information contact than that for Sorbian speakers. Translating a handbook of nano technology into Polish makes more sense than translating it into Sorbian or publishing Shakespeare's works in Frisian, another German minority language.

5. Background

To come back to the main issue of this volume: Language technology in Europe is not an extension of known technologies to new languages, but a multidimensional challenge for science, technology and politics of quite another order of magnitude. It will bind research groups, translators, software companies and politicians for the next 50 years at least.

There is a widespread conception that

- the rapid development of the Internet,
- with new web services,
- the globalisation of the markets and
- the increase of online transactions

are the main factors driving international research in language technology.

This argument is, at least in a European context, only partially valid. In the era when Internet was in its infancy, and most part of the online information was exclusively distributed in English, the Directorate General of EU "Linguistic Applications" was already concerned with the additional languages from the countries willing to join the European Union.

"With the expected enlargement of the EU following the accession of up to ten Central and Eastern European countries (referred to as CEECs, which is the usual EU abbreviation), the translation complexity takes a quantum leap. The current EU languages (n.R. situation in 1998) (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish) can be translated in 110 language combinations, as each of the 11 languages can be translated into 10 other languages. With the addition of 10 new languages (Estonian, Latvian, Lithuanian, Polish, Czech, Slovak, Hungarian, Slovenian, Romanian and Bulgarian) the complexity goes up to 21 x 20 = 420 language combinations, but there is no obvious political or linguistic justification for changing the European Union's official policy of supporting multilingualism, which finds its expression in the MLIS programme, among others."

(Poul Andersen, DG XIII EU Representative, in an article "Translation Tools for the CEEC Candidates for EU Membership - an Overview", *Terminologie et Traduction* 1.1998, pp.140-166).

One can only speculate about the development in Europe in multilingual language technology without the political changes of 1989. The rule-based machine translation system Systran was functional, with reasonable performance for the EU-languages and for the requirements of that time, namely translation of easy official documents.

We assume that the dramatic development of multilingual language technology in Europe was in fact driven by two forces: The new political context and the social impact of the Internet, rather than the economy.

We are also convinced that the European approach to multilingual language technology gave an impulse all around the globe to develop applications for various language communities: Recently systems for several Ethiopian languages appeared, a machine translation system for Quechua was presented (to quote only a few examples).

6. Overview

The present volume is structured in three sections that mirror the main challenges which researchers and developers have to face with the new languages from Central and South-Eastern Europe.

I. Linguistic Diversity

Some details were already illustrated by the previous paragraphs. We would only like to add here that the vast amount of complex linguistic phenomena of the new languages also imposed the "revival" of linguistic approaches and the development of hybrid approaches.

II. Language Resources and Tools (LRs)

Big bilingual/multilingual text collections and corresponding tools still form a huge gap, that languages in central and Southern/Eastern Europe had to overcome in a relatively short period of time. In an era where statistical approaches based on large training data sets were gaining more and more attention, the "new" languages were confronted with a big deficit. While today monolingual resources and tools (after almost 20 years and a dozen of European and nationally funded projects!) cover the minimal requirements of the BLARK-Kit (see http://www.blark.org/), the bilingual LRs are still lacking maturity and mass domain coverage.

A major drawback also was the lack of standardisation in the early 90ties. A lot of valuable resources were created for particular applications and in specific formats. Their re-use is still problematic. Thus, the current volume intentionally focuses on the development and use of generalised multilingual applications in globally accepted formats.

III. Multilingual Applications of Machine Translation and Cross-lingual Retrieval

This section presents system descriptions involving at least one language from Central or South-East Europe. The papers illustrate difficulties with generic frameworks like the statistical MT system Moses, when applied to lesser resourced languages. Even without any linguistic evaluation, the n-gram BLEU evaluation score reveals big differences among translation pairs ("462 Machine Translation Systems for Europe", Philip Koehn, Alexandra Birch, Ralf Steinberger, http://www.mt-archive.info/MTS-2009-Koehn-1.pdf). Still less investigated than machine translation is the conceptual cross-lingual retrieval and multilingual information fusion in extraction systems. We are proud to have even two contributions on these issues.

The present volume aims at giving an overview of extant resources and multlilingual applications for Central and South-Eastern European languages at a point when recent EU initiatives (e.g. META-NET and its subprojects) are starting for the first time to harmonise activities in language technology across Europe.

We are indebted to all authors who contributed to this volume and made its publication possible in relatively short time, an essential issue due to the fast development in the field.

We would like to thank very much the other members of the Editorial Board, Galia Angelova (Bulgarian Academy of Sciences, Bulgaria) and Joseph Mariani (CNRS-IMMI, France), who gave valuable feed-back to the contributors and editors, as well as John Judge from META-NET and Patrick McCrae from LangTech for their involvement in the editorial process.

CHAPTER ONE:

LINGUISTIC INSIGHTS

DATA CATEGORIES FOR MULTILINGUAL HARMONISATION-ORIENTED TERMINOLOGY WORK

ELENA CHIOCCHETTI

1. Introduction

This paper is based on the experience gathered within the LexALP[1] project, which aims to provide detailed analysis and subsequent harmonisation of the specialised terminology, in particular the legal terminology, used within Alpine Convention texts. The final objective of the comparative terminology work carried out within the project is to allow a group of experts, representing ministries, public institutions, universities and research centres of the entire Alpine arc[2], to determine translation equivalents in the four official languages of the Convention (one-to-one correspondence between the French, German, Italian and Slovene terms). Thus, the project intends to foster clear, efficient and, above all, consistent communication and text production for the future.

The Alpine Convention (AC) is a framework agreement signed in 1991 by all countries of the Alps (Austria, France, Germany, Italy, Liechtenstein, Monaco, Switzerland and Slovenia) and by the European Union for the protection and sustainable development of this mountain region. Besides the Frame Convention, nine Implementation Protocols were drafted since (Collectio 2004:7-13). With the exception of UN treaties, having four official language versions is uncommon for an international convention; usually there might be several translations, but

[1] LexALP is short for *Legal Language Harmonisation System for Environment and Spatial Planning in the Multilingual Alps*. The project is financed by the INTERREG IIIB "Alpine Space" programme and will be concluded at the end of February 2008. Information beyond the scope of this article can be found at the address http://www.eurac.edu/lexalp.

[2] A list of members of this "LexALP Harmonising Group" can be found on the Internet: http://217.199.4.152:8080/general/lexalp/index.php?id=members.

only few official versions. This makes it particularly challenging both to attain full equivalence and equal authenticity for all four texts as well as to provide a version that can be ratified without major problems in various national legal systems (Šarčević 1997:195-196). Even though some countries might share the same language, the legal system will differ; hence, the same term might be used to indicate different concepts (Sandrini 1999:16-17, de Groot 1999a:204 ff., de Groot 1999b:12 ff.). Apart from the European Union, which since 2004 has adopted all four official languages of the Alpine Convention, there are Austria, Germany, Liechtenstein and Switzerland having German as an official language, France, Monaco and Switzerland united by the French language and Italy and Switzerland, which both use the Italian language. This means that for all language versions, including Slovene (Slovenia and the EU), the text was ratified in at least two different legal systems. Given the nature of the legal languages, being always deeply rooted in their respective national traditions and realities (cf. Sandrini 1996:16, Sandrini 1999:10, de Groot 1999a:203, Gémar 1995:83), the need for clear, consistent and widely understandable terminology use within AC texts is paramount.

Furthermore, each Implementation Protocol was first drafted by different working groups (Alpensignale 2003:10) in the language of the respective chairing country and subsequently translated. The following states chaired the drafting process of the eight main Protocols[3]: Austria – Mountain Forests; France – Spatial Planning and Sustainable Development, Tourism; Germany – Nature and Landscape Protection, Soil Protection; Italy – Mountain Farming, Energy; Switzerland/Liechtenstein – Transport (Mayrhofer 19/06/2007). Due to this the quality of the texts can be very diverse in the different languages and this is especially true for Slovene, since no Protocol was originally written in that language.

2. Methodology

Given these premises, harmonising the terminology used within AC texts across the four languages becomes a daunting task. The need for harmonisation was recognised even after almost all Protocols had been declared harmonised from the point of view of language, style and content (Luzern 2000:5.6)[4]. The text of the Protocol of Decisions taken in the

[3] The ninth Protocol deals with the Composition of Controversies between the Parties.

[4] The Tourism Protocol and the Protocol on the Composition of Controversies are not included in the list.

Swiss town of Luzern does not state explicitly whether interlinguistic or intralinguistic harmonisation is meant, but according to Šarčević (1997:210) usually 'harmonisation' is referred to an *intra*linguistic process when dealing with international conventions. Still, many (quasi-)synonyms, variants, imprecise or incorrect translations survived in the AC texts.

The work methodology agreed on within the project to attain terminological harmonisation foresaw closely entwined steps (cf. Arntz 2002:119 ff.):

1. Definition and classification of basic thematic subfields and glossaries.
2. Extraction of all terminology used within the Frame Convention and its Protocols and their subdivision into the agreed on glossaries.
3. Description of the meaning and use of the concepts within the AC texts for all four official languages, whenever possible.
4. Comparison of the concepts found within AC texts with equivalent concepts in all legal systems under analysis, i.e. all national legal systems, the European law and international law[5] (cf. Arntz 1999:187 ff., de Groot 1999a:205 ff., de Groot 1999b:20 ff., Sandrini 1996:133 ff.).
5. Highlighting of legal, conceptual and linguistic differences that might cause a misunderstanding of the AC terminology within the legal systems of the Parties (cf. Sandrini 1999:32).
6. Preparation of harmonisation proposals to be evaluated by the group of experts.

This work intended to highlight which terms are more commonly used to designate the same (or in any case very similar) concepts within the above-mentioned legal systems, which terms were more widely understood, which terms actually had slightly or significantly different meanings. Legal and terminological comparison aimed not only at finding the 'best' term within the readily available linguistic forms, but also to propose appropriate translations where no adequate translation had been provided and, most of all, to clearly define several concepts. Thanks to this work, for many concepts used within AC texts now there are four harmonised equivalents designating it (one for every official language) and harmonised definitions. The AC definition might obviously differ in

[5] Even though Monaco and Liechtenstein are Parties to the Convention, due to lack of time and financial means the terminology used in these two countries could not be considered within the project.

some aspects from the national legal definitions given for the same terms; usually it is wider or more general. Thus, those definitions serve as explicit references and guidelines for all legal experts, translators, interpreters and drafters working with the AC.

Project results are made available freely through a dedicated terminological data base at http://www.eurac.edu/lexalp. The LexALP term bank allows terminologists seated in different Alpine countries to collaboratively work on the same term entries and publish the results directly online in real time. It was created within an existing lexicographic tool[6] that had to be adapted to suit concept-oriented legal harmonisation work. The main difference with other data bases ensuing from this fact is that for every legal system (and for every language within each multilingual legal system) monolingual terminology entries are created. These are then manually linked to a central node that represents the conceptual level. Conceptual relations are therefore expressed at this higher level and no cross references to any related term are found within the single term entries (cf. Sérasset et al. 2006).

As with any new terminological project, many initial efforts are put into defining the data categories or data elements needed within the term bank. For this purpose, there are dedicated ISO norms that help guiding the choices, such as, first and foremost ISO 12620. But as no standard can cover all possible types of needs, within each project the relevant data categories will have to be selected, defined, created and applied (Wright 2001:553-554). The following sections will deal with the challenges faced during the conception of the data base structure and the selection of data fields, given the specific structure and features of the LexALP term bank.

3. Data categories

3.1 Create project-specific data categories

The LexALP term base is conceived for a specific type of prescriptive terminology work, i.e. harmonisation-oriented comparative terminology work. Next to the types of data categories commonly present in most term banks (term, definition, context, source, etc.) a specific data category had to be foreseen in order to indicate the harmonisation status of the terms, as determined during the final decision process by the group of experts. It is therefore necessary to be able to mark some terms as officially validated or rejected. To this purpose, the data category ‘normative authorisation’ in

[6] For more information see http://www.papillon-dictionary.org.

the ISO 12620 served as a perfect model for the data category that was then applied within the LexALP term bank. The former is defined as *a term status qualifier assigned by an authoritative body, such as a standards body or a governmental entity with a regulatory function.* The group of experts validating the terminology within the LexALP project is not an official body with strong regulatory powers. Still, within the frame of the project their opinion can definitely be considered as authoritative. Hence, the data category 'harmonising status' was created with a pick-list of three items, i.e. 'harmonised', 'rejected' and 'unknown'. The last item serves for those AC entries that have not yet undergone or will not undergo the harmonisation process.

3.2 Adapt existing data categories

For each AC concept analysed within the project there are several entries pertaining to the different terms that express it at AC level and to the terms designating equivalent concepts in other legal systems. So, in order to label the entries as clearly belonging to different legal realities it was necessary to mark them accordingly. The most obvious solution would have been introducing a label called 'legal system'. However, some of the subfields under analysis are not regulated at national level in some countries, but at a lower legislative level (*Land*, region, canton) and must therefore be terminologically described at that level[7]. In Austria the sub-national level consists of Federal States, i.e. the *Bundesländer*, whereas Italy has administrative subdivisions called *regioni*[8]. The latter cannot be considered separate legal systems within the Italian system. Hence, it was difficult to find a common label for this data category. To avoid this problem, a good solution was found in the ISO 12620 norm, which gives a

[7] For example, in Italy the protection of the environment is regulated by the State, whereas hunting and fisheries, forests, tourism or agriculture are exclusively of regional competence (cf. art. 117 of the Italian Constitution). Also, whereas landscape protection issues are regulated at the level of the single *Länder* in Austria, in France such strong delegation of competences is not implemented (cf. also Chiocchetti & Lyding 2006:509).

[8] Austria is a federal state of nine *Bundesländer*, whereas Italy is a unitary state subdivided into 20 administrative divisions, the local bodies called *regioni*. Each Austrian *Bundesland* has its own constitution and exclusive competences, even though a quite high number of subject fields are of federal competence (cf. § 10-11 of the Austrian *Bundes-Verfassungsgesetz*). Austria is therefore often considered as a rather 'centralist' federal state. As concerns Italy, the reform of article 117 of the Italian constitution did indeed confer a wider autonomy to the regions, but the unity of the state remained untouched.

definition of the data category 'geographical usage': *Term usage reflecting regional differences. […] [T]he content […] should be a country code as specified in ISO 3166 […]*.

Actually, as the examples given in the ISO norm show, this data category usually serves to distinguish between regional variants such as 'lorry' (British English) versus 'truck' (American English). However, within the LexALP project, the use of country codes and country flags allowed this data category to apply to the distinction between term entries of different legal systems. This is particularly useful in cases where the same term is described from the viewpoint of different legal systems. For example, the Austrian concept of *Naturpark* encompasses not only the conservation of landscape, nature and species as it does in Germany: information and education measures for the public must also be foreseen; otherwise it cannot become a *Naturpark.*[9] In such cases not only the contexts but also definitions will differ from each other for the two terms. As a consequence, it is fundamental to be able to keep the two entries clearly separate. Currently, the data category 'geographical usage' is used to indicate the legal frame of reference for each single monolingual term entry and has proven an intuitive, user-friendly solution.

3.3 Define permissible instances of data categories

When defining a term bank structure it is not just the selection of relevant data categories that must be carefully considered, also the items that will be available within all closed data categories must be determined. Within the LexALP project deciding which subfields and themes were to be dealt with was not a straightforward task. First of all, the Permanent Secretariat of the Alpine Convention strongly suggested basing terminology work on article 9 of the Protocol 'Spatial Planning and Sustainable Development'. It lists the main subject areas to be considered for sustainable development and includes (1) regional development policy, (2) rural areas, (3) urban areas, (4) nature and landscape protection and (5) transport. Unfortunately, this subdivision does not look like any structure known at national or European level and had to be adapted, integrated and further specified by different national legal experts. The result was to be a list of smaller subfields according to which the terminology of several different legal systems could be accommodated. Suffice it to say that for the Italian legal experts no subdivision of 'transport' could be considered

[9] See for example § 25 of the Burgenländisches Naturschutz- und Landschaftspflegegesetz and § 27 of the German Bundesnaturschutzgesetz.

complete without the glossary 'sea transport', which obviously is of very limited interest both for the inner Alpine countries and for the AC itself. Some compromise had to be found especially in the realm of urban areas, so as to respect both the peculiarities of the AC and of the national legal traditions. Even more so because in this matter the national categorisations and subdivisions diverge so much as to make them almost impossible to compare. Finally, after referring also to the European level, the structure shown below in figure 1 was developed. It presently serves both for the classification of all terms in the terminology data base and at the same time for the classification of all texts collected in the LexALP corpus.[10]

3.4 Apply data categories

Past experiences with legal comparative terminology work of the project partners showed the need to sometimes allow inserting information going beyond the definition of a concept. This helps to highlight conceptual differences between different systems properly. It is especially true for classical terminological definitions stating only the *genus proximum* and the *differentiae specificae,* because they might be sufficient to clearly place a term within the conceptual system of one legal system, but too succinct to explain all differences existing between concepts belonging to different legal traditions. This is due to the fact that legal concepts are socio-cultural products which are strongly dependent on their legal system of reference: full equivalence is rarely to be encountered when comparing different legal systems (Arntz 1993:6, Sandrini 1996:138, Šarčević 1997:232). Hence, an open data category 'note' was foreseen in the entry structure of the LexALP term bank. It was meant to accommodate first and foremost those comparative notes that were necessary to ease the harmonisation process of AC terms.

[10] The LexALP corpus consists of over 3000 documents on environmental law, spatial planning and sustainable development of all legal systems under analysis. The texts were selected by legal experts and classified according to the above-mentioned structure. Total corpus size currently amounts to about 20 million words (cf. Chiocchetti & Lyding 2006).

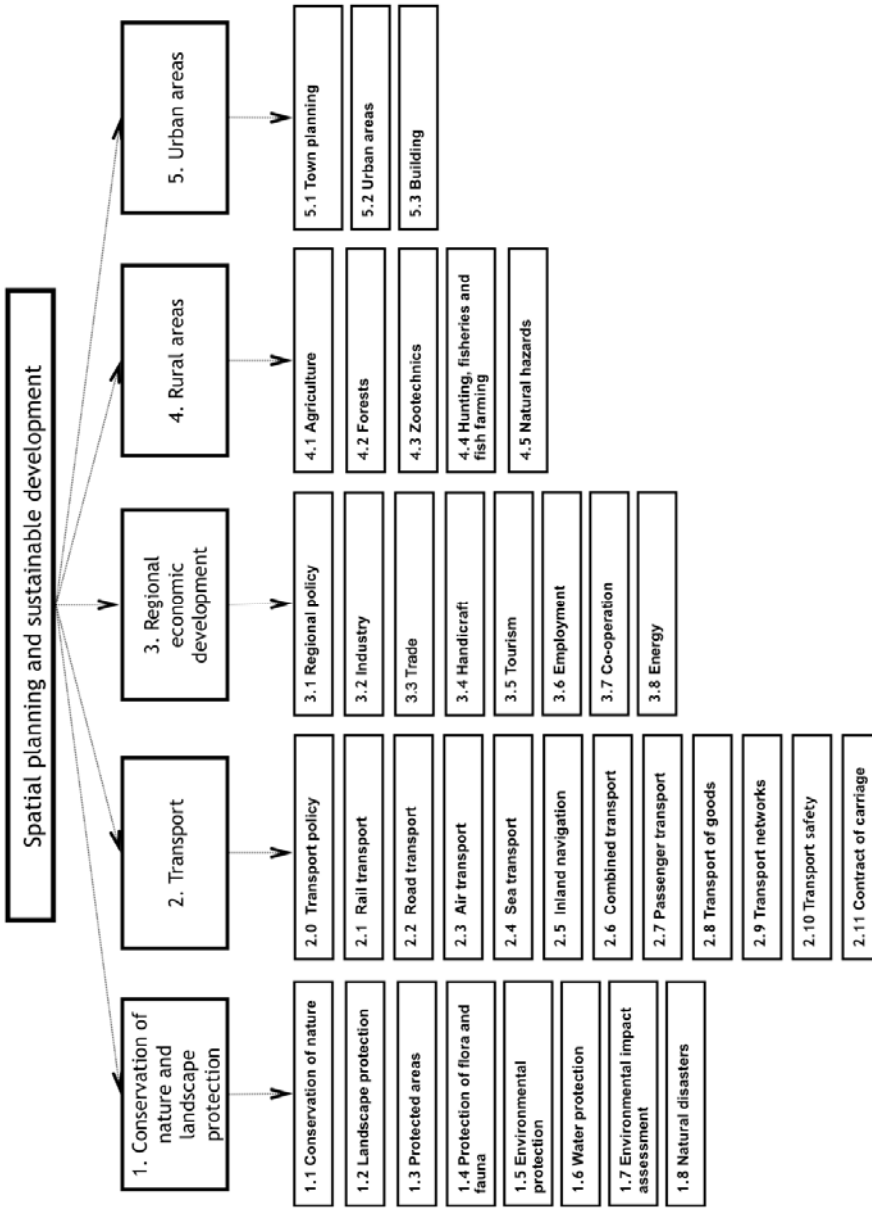

Figure 1: instances of the data category ‘subject field’

In the ISO 12620 norm the data category 'note'[11] is not underspecified. It is described as *supplemental information pertaining to any […] element in the data collection, regardless whether it is a term, term-related, descriptive, or administrative*. Two more specific data categories are suggested for use whenever possible, namely 'usage note' and 'transfer comment'. As regards the first, (*a note containing information on the usage of the associated term*) its subcategory 'geographical usage' is used in the LexALP term bank as explained in section 3.2. Also the data category 'frequency' is part of the data model. The second category, 'transfer comment' (*note included in a term entry providing more explicit information on the degree of equivalence, directionality or other special features affecting equivalence between a term in one language and another term in the same or a second language*) was originally not foreseen, because notes were supposed to be written mainly for terminology entries concerning the AC and with the goal of facilitating the harmonisation process rather than the translation process. Furthermore, analysing the different degrees of equivalence between all involved legal systems singularly would have gone beyond the aim of the project.

During practical terminology work over four languages and nine legal systems[12] it was realised that the use of the note field had been quite diverse in practice, notwithstanding the clear usage indications. At least seven different types of information where inserted into the data field 'note', according to the specific situation/problem the terminologists encountered:

[11] Admitted names for this data category are also 'remark' and 'comment' (ISO 12620).

[12] The Alpine Convention is treated as the ninth legal system in the term bank, even though, strictly speaking, it does not constitute a legal system *per se*. It being an international convention, the AC would have to be considered part of international law. However, in order to be able to clearly distinguish its terminology from the terminology of other legal systems and to give the users clear indications on the harmonised/rejected terms, the AC entries have been kept separate.

- **legal notes**: legal information, beyond the definition, referred to one legal system, e.g.

<Term> Verbandsklage
[…]
<Geographical usage> AT
<Domain> 1.1 Naturschutz
<Definition> Klage, bei der bestimmte Vertretungen von allgemein als gerechtfertigt anerkannten Interessen zur Entscheidung einer Rechtsfrage des materiellen Rechts Antragsrecht haben, um Ansprüche ihrer Mitglieder durchzusetzen oder Testprozesse ohne das Prozeß- und Kostenrisiko einer Einzelrechtsverfolgung führen zu können.[Document : VfGH, 11.12.1996, G52/95/Voltmer]
<Source> VfGH, 11.12.1996, G52/95/Voltmer
[…]
<Note> Verbandsklagen können etwa nach § 14 UWG, §§ 28 ff. KschG und § 54 Abs. 2 ASGG, § 6 Abs. 3 GleichbehandlungsG, § 12 Abs. 1 RabattG, § 7 Abs. 2 NahversorgungsG 1977, § 44 Abs. 1 KartellG 1988 und § 14 UWG erhoben werden.
<Source> VfGH, 11.12.1996, G52/95/Voltmer

- **comparative notes**: contrastive comments referred to two or more legal systems. For example, the note below explains that the term used in the German version of the AC Protocol on Nature and Landscape Protection has a different meaning than the Italian term used in the same paragraph (*acquisto di esemplari*):

<term> Kauf von Exemplaren
[…]
<Geographical usage> AC
<Domain> 1.4 Schutz der Fauna und der Flora
<Domain> 3.3 Handel
<Definition> Vertrag, durch den sich der eine Teil (Käufer) zur Zahlung eines vereinbarten Kaufpreises verpflichtet und der andere Teil (Verkäufer) zur endgültigen Übertragung von lebenden oder toten Tieren oder Pflanzen, von Teilen und Derivaten davon oder daraus gewonnenen Produkten.
<Source> KÖBL 2001: 269/RL 92/43, Art. 1, m/Chiocchetti
[…]
<Note> Die Termini „Kauf“ und „Verkauf“ (von Exemplaren) beziehen sich im deutschen Recht beide auf einen gegenseitigen Kaufvertrag. Auf Italienisch kann man „Kaufvertrag“ nur mit „compravendita“ [Kaufvertrag] oder „vendita“ [Verkauf] wiedergeben. „Acquisto“ bezieht sich hingegen auf die Tätigkeit, mit der man entweder das Eigentum oder den Besitz einer Sache (z.B. eines Exemplars), eines Rechts usw. erwirbt.
<Source> Chiocchetti

- **linguistic notes**: purely linguistic information (e.g. frequency according to text type, singular vs. plural usage etc.), e.g.

  <Term> Landschaftselement
  [...]
  <Geographical usage> DE
  <Domain> 1.2 Landschaftspflege
  <Definition> Bestandteil einer Landschaft. Dazu gehören Böden, Gesteine, Klima, Wasser, Vegetation, Fauna und die Maßnahmen des Menschen in der Landschaft.
  <Source> UWLX 2000:689
  [...]
  <Note> Der Terminus wird hauptsächlich im Plural verwendet.
  <**Source>** Scanzoni

- **terminological notes**: notes on conceptual differences between terms (e.g. conceptual discrepancies such as broader/narrower meaning, etc.), e.g.

  <Term> habitat
  [...]
  <Geographical usage> AC
  <Domain> 1.3 Aree protette
  <Definition> [L]uogo o tipo di sito dove un organismo o una popolazione esistono allo stato naturale.
  <Source> CBD 1992, art. 2
  [...]
  <Note> Il termine viene a volte impiegato anche per indicare l'ambiente in cui vive una comunità di organismi e, in questo caso, diventa virtualmente sinonimo di biotopo.
  <Source> GaMa 1995:359

- **encyclopaedic notes**: any other information on the concept that might seem useful (not necessarily legal information), e.g.

  <Term> rete nazionale e transfrontaliera di aree protette
  [...]
  <Geographical usage> AC
  <Domain> 1.3 Aree protette
  <Definition> Rete ecologica costituita da aree protette transfrontaliere, aree protette di grandi dimensioni e consorzi di aree protette.
  <Source> RETR 2004:119/Rampino
  [...]
  <Note> Nell'ambito dello studio [mandato Convenzione delle Alpi: "Aree protette transfrontaliere e rete ecologica nelle Alpi"], sono state registrate,