Methods, Models and Standardization Issues for the Creation of Linguistic Resources: the Case of Under-Represented Languages

Claudia Soria, Monica Monachini

ILC-CNR
Via Moruzzi 1, Pisa, Italy
{claudia.soria, monica.monachini}@ilc.cnr.it

Keywords: less widely available languages, multilingual terminological resources, resource production methodology, and standards.

Abstract Availability of Linguistic Resources for the development of Human Language Technology applications is nowadays recognized as a critical issue with both political and economic impact and implications on the sphere of cultural identity. Many languages have little or no information technology available. This paper reports about the experience gained during the *INTERA* European project for the production of multilingual terminological lexicons for those languages that suffer from poor representation over the net and from scarce computational resources, but yet are requested by the market. It discusses the procedure followed within the project, focuses on the problems faced which had an impact on the initial goals, presents the necessary modifications that resulted from these problems, evaluates the market needs as attested by various surveys, and describes the methodology that is proposed for the efficient production of Multilingual Terminological Lexicons.

1 Introduction

Language Resources are central components for the development of Human Language Technology applications. The availability of adequate Linguistic Resources for as many languages as possible plays a critical role in view of the development of a truly multilingual information society. It is no mystery that the task of producing language resources is an extremely long and expensive process. For most western languages, and English in particular, however, this is partly softened by large and often free data availability, good representativeness, and significant size, together with availability of language processing tools. There is plenty of languages, however, for which this picture is far from being adequate. Many languages actually suffer from poor representation and scarcity of raw material, not to mention the availability of robust processing tools. It is imperative to try to reach a balance of language coverage in order to avoid a *two-speed Europe* (Maegaard *et al.* 2003). This awareness gave rise to coordinated efforts, both at national and European level, in the direction of reducing this gap (Calzolari *et al.* 2004) and enabling less-favoured languages with respect to language technology. The concepts of BLARK, i.e. the definition and adoption of a standard Basic Language Resource Kit for all languages, a minimal set of language resources necessary to the development of language and speech technology (Krauwer 1998), goes exactly in this direction.

In *INTERA*, the expression "less widely available languages" has to be interpreted in the sense of (Gavrilidou *et al.* 2003, Gavrilidou *et al.* 2004), that is, of "less widely available in the digital world". This concept has been developed in response to a survey, also conducted in the framework of the *INTERA* Project, aimed at the identification of users' needs and expectations concerning language resources. Although western European languages have been confirmed as having the highest amount of request, the survey clearly demonstrates that there is an increase in demand for Balkan and eastern European languages. Under this respect, it has to be noted that the notion of "less widely available languages" is by no means to be interpreted as a synonym to "less widely spoken languages"; in fact, many of the languages to which the former concept seems to apply, such as eastern European languages and Balkan ones, actually appear among the forty most widely spoken languages all over the world (http://www.globallanguages.com). For instance, Russian, Polish, Ukrainian, Romanian, and Serbo-Croatian appear in the 8th, 22nd, 23rd, 32nd, 37th rank respectively. Nevertheless, these languages still seem to be rather under-represented as regards digital content, although there is an increase in Balkan LRs production - tendency encouraged by national and EU activities as well - as attested by surveys conducted in the framework of other European projects (i.e. ENABLER, www.enabler-network.org). Despite their limited availability, Balkan and eastern European languages are highly requested by the digital content market, thus making the issue of resource production even more crucial.

The *INTERA* European Project had a twofold task: on the one hand it was aimed at producing multilingual parallel corpora and terminological lexica for some languages that were identified as belonging to the class of "less available ones", i.e. Greek, Serbian, Bulgarian and Slovene. On the other hand, another aim of the project was establishing a reference production model for multilingual resources, which is up-to-date, compliant with existing standards and yet viable and attractive for digital content producers. In this paper we report about the experience gained during the *INTERA* Project for the production of a model for multilingual terminological lexicons.

The particular point of view of trying to design a production model for multilingual terminological lexicons inevitably brought us to try to conciliate two different, opposite forces. On the one hand, there are the needs and requirements expressed by users of the digital content market, as emerged through a thorough examination of eContent professionals' practices and policies as for LRs. From this point of view, any prospective resource should aim at completely satisfying user needs and requirements, as well as complying with existing standards. On the other hand, there is evidence of unavoidable production gaps and the users' documented desiderata are in conflict with the actual viability and feasibility of language resources, as is documented by various surveys of language resources (Gavrilidou and Desipri 2003; Gavrilidou *et al.* 2003).

This means that a realistic production model should also take into consideration very basic problems of data availability, representativeness, and size, together with availability of language processing tools. A realistic model for the production of multilingual terminological lexicon, thus, rather than describing an ideal situation, which would be far from reality, should consider a variety of possible situations thus anticipating possible shortcomings in all stages of production.

The paper is organized as follows: Section 2 describes user needs; Section 3 illustrates possible scenarios to be faced as for data availability, format, and annotation, whereas Section 4 presents the actual scenario we were confronted with. Section 5 presents the methodologies and techniques used for the construction of the *INTERA* multilingual terminological resource. Finally, Section 6 presents the conclusions.

2 User Needs

According to the requirements emerged from *INTERA*'s user needs survey and from the outcomes of relevant past initiatives, a multilingual terminological lexicon should fulfil both resource-related and content-related requirements. Resource-related requirements are those referring to overall characteristics such as use of common, widely accepted and widely used standards, availability of readily accessible information, and validation. Content-related requirements, on the other hand, refer to the availability of particular types of information in the terminological entries, which users assess as essential, desirable, or irrelevant. Equivalents in other languages and grammatical information about a given terminological entry are considered as essential information, while definitions, conceptual relations, a domain code indication, and reliability codes are seen as desirable.

3 Foreseen Scenarios

The quality of the final resources produced is strictly intertwined with the existing source material, its (re)usability, the presence of linguistic annotation or the availability of tools to perform linguistic analysis, the compatibility/reusability of different linguistic analyses, etc. Data availability, representativeness, and size, together with availability of language processing tools are crucial factors to be taken into account in a very realistic production model, since they are variables with strong repercussions on the task and the process of corpus production and terminology extraction. Different possible scenarios can thus be foreseen for the production of multilingual resources, all of them having an impact over the task of term extraction. The quality of the multilingual resources dramatically differs depending on the scenario, i.e. according to whether a corpus is *parallel* or *comparable*, whether there is a unique pivot language available or not, etc. The possible scenarios envisaged can be summarized as follows:

In the "ideal" scenario, the extraction task can be performed working on parallel specialized texts with a pivot language (hopefully English) for which NLP tools and resources are available. It can be expected that sufficiently "clean" lists of candidate terms can be extracted that are to be confirmed by the terminologists. In this situation, we can aim at a *truly* multilingual database, where the lexicon will be the same across languages and terms will be all interconnected and corresponding to each other.

In the "worst" solution, where there are no truly parallel texts but sets of parallel texts without any pivot language, we only can resort on statistical procedures of term recognition. The greatest risk is to produce a list of candidate terms with possible *noise* or *silence*[1] and where human involvement is massive.

In a "mid-way" solution where sets of pairs of corpora are available, the resulting terminological lexicon is not a truly multilingual homogeneous one, but a set of terminological lexicons in different languages where terms are likely not to be the same throughout the lexicons.

4 Deviation from initial goals: The Actual Scenario

The actual configuration of parallel corpora found in *INTERA* was actually different from expected. The scenario we were confronted with was rather similar to the above mid-way solution: the final collection is represented by a *comparable multilingual corpus* as opposed to a parallel multilingual corpus. Instead of having the same texts in all languages, there were pairs of different texts loosely belonging to the same domains (namely, *law, tourism, health, and education*). For each pair, English always represents one member (the *pivot* language), while the other is Greek, Bulgarian, Slovene, or Serbian.

The Table below illustrates this situation in detail.

Domain	**Languages**			
	Greek	**Bulgarian**	**Serbian**	**Slovene**
Law	x	x	x	x
Health	x		x	
Education	x	x	x	
Tourism	x			
Environment	x			
Finance			x	
Politics		x		

Figure 1 : Distribution of languages and domains

The first consideration to be made concerns the *degree of multilingualism* of the terminology: since the texts are not truly parallel, the final terminology is not a truly multilingual one. In other words, the lexicon is not the same across languages and terms are not all interconnected and corresponding to each other. Instead, for each English–*X* pair we derived the

[1] Noise and silence are commonly used in terminology as complementary of precision and recall respectively.

corresponding terminology, thus arriving at a bilingual (English-language–*X*) terminology for each domain. Since the domains are at least partially overlapping, some terms occurring in one terminology also occur in another one, thus enabling to build truly multilingual links at least for a subset of terms. Given the situation illustrated in the Table above, the only domain for which a quadri-lingual partial terminology is feasible is the Law domain. The Education domain yields a tri-lingual terminology, and the health domain a bi-lingual one. The Tourism, Environment, Finance, and Politics domains are monolingual terminologies.

The second consideration is related to the range of technical solutions adopted for automatic term extraction. The availability of the same pivot-language for all target languages proved useful, especially because the target languages are under-represented ones, for which few reference corpora and NLP tools are available. On the contrary, there is a huge amount of resources (corpora, lexica and tools) available for the English language, and this allowed us to opt for a combination of statistical and NLP procedures, as illustrated in more detail in the next Section.

5 Terminology Extraction

Terminology can be considered the surface realization of relevant domain concepts (Cabré, 1992; Sager, 1990). Candidate terminological expressions are identified either by hand, or in a semi-automatic manner. Semi-automatic procedures for terminology extraction usually consist in shallow techniques that range from stochastic methods to more sophisticated syntactic approaches (Jacquemin, 2001; Bourigault, Jacquemin, L'Homme, 2001).

All of them, however, converge in identifying terms mostly on statistical grounds, on the basis of its relative frequency in a corpus, possibly augmenting this measures with filters capturing the domain specificity of a term. Although not theoretically correct (as the status of "termhood" is in principle independent on the number of occurrences, and a *hapax* might well be a term), this practice is rooted in computerized terminology, where computer-aided text analysis and the possibility of processing large amount of information have changed the bases of terminology compilation, as well as how the appropriateness of terms is conceived and the degree of human intervention in the whole process. In this particular context, we adopted a hybrid approach to terminology extraction from multilingual parallel texts, combining statistical and symbolic techniques.

5.1 The data

As introduced above, the data available for the task of automatic term extraction come under the form of four parallel corpora: English-Greek, English-Serbian, English-Slovene and English-Bulgarian. Each parallel corpus is further organized according to the particular domain to which the texts of the corpora belong: while the English-Slovene corpus covers the law domain only, the English-Greek corpus, for instance, covers as many as five domains, i.e. law, education, health, tourism, and environment.

The size of available data is important for determining the coverage of the terminological resource, since more data mean more terms. However, it is important also for the quality of the terminological resource, as the automatic procedure needs an amount of data reaching a level of statistical relevance to yield high-quality data. Unfortunately, the available data dramatically differed in size both across the different domains and across the different languages. The biggest data were available for Greek and Serbian (59 and 69 Mb respectively), while Bulgarian and Slovene were represented globally only with 24 and 33 Mb.

The richest domain is represented by law (129 Mb), followed at a distance by education (20 Mb) and health (14 Mb). This difference among domains has an obvious consequence over the overall amount of terms that can be made available as a result of the extraction process. In other words, there will be domain-specific terminologies that will be very different in size and hence term coverage. This situation is clearly depicted by the case of the terminology for the health domain. The corpus data amount to 13 Mb for Greek and 1Mb for Serbian. The highest quantity for Greek allows to extract more candidate English terms, as easily foreseen, but, most importantly, to produce less candidate translators and of better quality: while for Greek the candidate translators/terms ratio is of 1,5, for Serbian it is of 4,1.

	Candidate terms	**Terms**	**Candidate translators**	**Translators**
Serbian	734	488	2012	201
Greek	1710	1052	1580	826

Figure 2: Candidate translators/terms ratio for Greek vs. Serbian

5.2 Extraction procedure

The task of automatic term extraction was organized around three main phases:

1. Automatic extraction of terms from the English components of the parallel corpora. The English language is henceforth defined as the "pivot language";
2. Automatic identification of candidate translators in the target languages;
3. Manual verification of the candidate translators found with the automatic procedure.

5.2.1 Extraction of English candidate terms

The objective of the first step is the identification of terms for a given sub-language; it is assumed that these terms should represent those that most probably are peculiar for a specific sub-domain. Under this assumption, the terms that will be identified will represent the candidate terms for a specialised (domain-specific) lexicon.

Candidate single terms are extracted by comparing the relative frequency of lemmas inside each domain and language specific subcorpus against a lemma-based frequency lexicon of the *British National Corpus*, which was used as a reference corpus.

In more detail, the comparison between the frequency distributions of terms in the general lexicon and that of the different domain-specific lexicons was performed adopting a mathematical function evaluating "the distance of the frequency of domain-specific terms from the frequency which was expected on the basis of the general lexicon".

We compared the lists generated adopting several different mathematical formulae, among which are the following:

$$d1 = f_{r\,(\text{specialized lexicon})} - f_{r\,(\text{general lexicon})}$$

$$d2 = f_{r\,(\text{specialized lexicon})} / f_{r\,(\text{general lexicon})}$$

$$d3 = \log(f_{r\,(\text{specialized lexicon})} / f_{r\,(\text{general lexicon})})$$

where f_r represents the relative frequency of a term inside the lexicon.

Terms, however, are not simply represented by single terms. Examples include compounds (*credit card*), adjective-noun (*administrative procedure*) or complex noun phrases (*principle of equal treatment*). We thus specified a bunch of basic syntactic rules expressing constraints over syntactic patterns in order to select candidate multi-word terms.

In order to avoid over-generation problems, some corrective measures have been applied, most notably by specifying either lists of words to be discarded *a priori* (stop-word lists) or different values of the threshold under which a candidate is automatically rejected. The threshold is each time adjusted depending on the overall size of the parallel corpora under analysis and empirical measures.

5.2.2 Extraction of candidate translators

Once candidate terms are identified for English, we turn to the task of automatic identification of candidate translators in the target languages. To this end we exploited the structuring information available in the parallel corpora from which the terminology was to be extracted. Since the sentences in the target language texts are aligned to those of the pivot language, it is easy to select a suitable search space for any candidate term. The algorithm for the extraction of candidate translators consists of the following steps:

1. Selection of the *source region set* from the pivot language corpus;
2. Extraction of *target region set* from the target language corpus;
3. Search Extraction of lemmas from target region set;
4. Ordering of the lemmas contained in the search target region set according to a *ranking function*;
5. Selection of candidates.

Given a candidate term *t* (in English), the target region set inside the target language corpus is easily identified thanks to the parallel structure of files to be processed: each region of the

English corpus containing the term *t* is uniquely associated with a region of the target language corpus.

Then, the lemmas from the target region set are extracted, filtering out lemmas belonging to “non significant grammatical categories” (e.g. conjunctions, prepositions).

It was observed that the target language lemmas could be classified on the basis of their “probability” of being a translation of a given term by means of simple frequency analyses.

This classification is obtained through the synthesis of a ranking function. Several hypotheses were considered, all of them aiming at highlighting the statistical “idiosyncrasies” of the translating lemma.

The best performing measure is the following:

$f(l)= r(l)-q(l)*|I|$

Where $r(l)$ is the number of regions of the target region set containing at least one occurrence of lemma l, $q(l)$ is the ratio between the number of regions containing lemma l and the total number of regions in the corpus; $|I|$ is the total number of regions of the target region set.

5.2.3 Validation and production of terminological entries

The lists of candidate English terms and their corresponding candidate translations in the other languages (lists of single word terms and multiword terms as produced by the tool) were presented to human validators, who were all native speakers of the selected languages. The validators' task consisted in examining the lists and marking bilingual pairs of terms (English – their mother tongue) as correct, based on the following criteria:

(a) the pair is indeed a term of the specific domain (and not a general vocabulary word) AND

(b) the translational equivalence is also correct.

The terms identified as correct pairs were further lemmatized (single word terms and multiword terms), and the final lists produced by the validators were suited for the production of the multilingual terminological entries.

Since compliance to a reliable standard framework is a pre-requisite for ensuring sharing, reusability and exchangeability of data, the TMF family of formats was taken as the reference model to encode the *INTERA* terminological entries. TMF stands for Terminological Mark-up Framework (ISO 16642 2001), an international standard designed in the framework of the ISO initiatives to support the creation and use of computer applications for terminological data and exchange of such data between different applications. Being a meta-model for terminology mark-up, TMF allows for the specification of user-defined mark-up languages (called TMLs). A TML makes it possible to design the encoding format of a terminological collection according to specific needs. In designing the *INTERA* TML we tried to harmonize users’ needs with the realistic considerations when dealing with under-represented languages.

6 Conclusions

The approach to multilingual terminology lexicons adopted and described in this paper cannot be seen as a standard practice nor is it to be considered as a recommended practice in terminology building. In fact, there is no such practice for all purposes. There only are better solutions under certain conditions. We claim, however, that the procedure adopted is a viable and fruitful one given the following conditions:

- Data are sparse
- No NLP tools are available for the target languages
- No reference corpora are available for the target languages but many NLP tools and reference corpora are available for one language (the *pivot*)

Moreover, this experience taught us some lessons about the more general task of building terminological resources for languages suffering from scarcity of widely available data and processing tools. Thus, besides the actual production of the resources, a parallel result has been the identification of gaps and shortcomings in the process usually employed by LRs producers (or users who might wish to create their own LRs) and to suggest ways of remedying them.

At a general level, the production methodology is heavily influenced by the following factors:

- Lack of integration among computer tools working at different levels of analysis.
- Lack of compatibility among the resources themselves. This means, for instance, not only enforcing compatibility in data encoding and representation, but also ensuring that the resources are compatible from the point of view of the additional, linguistic and non-linguistic information, which is added to the raw data. Once again, compliance with agreed-upon standards is recommended, as well as harmonization among the different tag sets used in the various resources. Ideally, all resources should use the same convention of linguistic annotation; when this is not possible, it is recommended that a harmonized tag set is used, or that conversion procedures from the proprietary tag set to a common, standardized one is provided.
- The limited number of existing corpora, especially in languages other than English.
- The particular configuration of resources available. The particular methodology to be adopted for the production of multilingual terminological resources must be carefully adjusted to the idiosyncratic situation to be handled, where by situation we mean the type of languages, the quantity and quality of resources, and the purposes for which the resource is being built.

In conclusion, the experience gained during the *INTERA* project calls for more resources, of good quality, and compliant with sound standards. Lesser-favoured languages can benefit from building parallel resources where English represents one language. Another general recommendation is that a criterion of *practical feasibility* be followed, thus balancing the constraints imposed by corpus size, languages, and users' and standards requirements. This is deemed the only viable and reasonable solution, especially from the point of view of prospective users that will have to apply the model that is the outcome of the *INTERA* project.

References

BOURIGAULT D., JACQUEMIN C., L'HOMME M.-C. (EDS) (2001), *Recent Advances in Computational Terminology*, Amsterdam & Philadelphia, John Benjamins.

CABRÉ, M.T. (1992), *Terminology. Theory, methods and applications*, Amsterdam & Philadelphia, John Benjamins.

CALZOLARI N., CHOUKRI K., GAVRILIDOU M., MAEGAARD B., BARONI P., FERSØE H., LENCI A, MAPELLI V., MONACHINI M., PIPERIDIS S., (2004), *ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs,* in LREC-2004 Proceedings, Lisbon.

GAVRILIDOU M., DESIPRI E. (2003), *Final Version of the Survey*, ENABLER Deliverable 2.1.

GAVRILIDOU M., DESIPRI E., LABROPOLOU P., PIPERIDIS S., MONACHINI M., SORIA C. (2003), *Technical specifications for the selection and encoding of multilingual resources,* INTERA Deliverable D5.1.

GAVRILIDOU M., GIOULI V., DESIPRI E., LABROPOLOU P., MONACHINI M., SORIA C., PICCHI E., RUFFOLO P., SASSOLINI E. (2004), *Report on the multilingual resources production,* INTERA Deliverable D5.2.

KRAUWER S. (1998), *ELSNET and ELRA: A common past and a common future,* in *ELRA Newsletter,* Vol.3, N. 2.

MAEGAARD B., CHOUKRI K., MAPELLI V., NIKKOU M., POVLSEN C. (2003), *Language Resource – Industrial Needs,* ENABLER Deliverable D4.2, Copenhagen.

JACQUEMIN, C. (2001), *Spotting and Discovering Terms through Natural Language Processing*, Cambridge, MA and London, The MIT Press.

SAGER, J.C. (1990), *A Practical Course in Terminology Processing*, Amsterdam & Philadelphia, John Benjamins.