# Dealing with Large Corpora for Ontology Population

**Yuliya Korenchuk (1,2)**
(1) LiLPa (Linguistique, Langues, Parole), EA 1339, Universit de Strasbourg
(2) Rebuz SAS, Strasbourg
yuliya.korenchuk@yahoo.fr

## 1 Introduction

Multilingual ontology population from texts, i.e. addition of new terms in an ontology, requires a suitable parallel or comparable corpus. In this paper, we aim to check whether the corpus selected for our project suits the ontology we want to populate. The corpus for ontology population should not only reflect a specific domain and have a sufficient volume of data, as discussed in (Delpech et al., 2012), but also suit the initial ontology. Using an existing corpus can be an efficient solution used in many projects (Cimiano, 2006; Bouamor, 2014; Pinnis, 2014). However this option is less reliable in the case of a large multi-domain corpus and an ontology which might not cover all the domain concepts. The need for suitability between text corpora and ontology is expressed by (Aussenac-Gilles et al., 2006) who underlined the importance of text type in the corpus, the ontology application, the validation criteria and set up. The text layout can also play an important role: some projects aim to use extralinguistic information for ontology population (Kamel et al., 2013), while others concentrate on the comprehensiveness of the text (Faber et al., 2006).

In this case study, we set up an experiment checking whether a corpus is suitable for ontology population, based on the example of the large parallel (English, French and German) corpus PatTR[1] (Wäschle and Riezler, 2012) and the EcoLexicon[2] terminology knowledge base which we use in our project.

## 2 Resources

### 2.1 Corpus

The PatTR corpus is a large[3] collection of parallel segments from patents organized by language pairs. These segments are classified into files according to their position in the patent structure (title, abstract, description or claims) (Wäschle and Riezler, 2012). All the language pairs have their metadata files which contain essential information (the IPC[4] code, the reference, etc.) for each segment. As the different domains are mixed, the metadata play a crucial role for our project.

### 2.2 Ontology

The terminological knowledge base EcoLexicon is developped by the LexiCon research group at the University of Granada. The resource is designed according to the principles of Frame Based Terminology (Faber et al., 2005; Faber et al., 2006; Faber et al., 2009; Faber, 2011; Araúz et al., 2011). It contains 3,547 concepts and 19,712 terms (cf Table 1) on the topic of environment in seven languages, including English, German and French. The terms are connected by generic-specific, part-whole and non-hierarchical relations. The latter refer to the behaviour of the concepts in a domain-specific or a general semantic frame (Faber et al., 2009).

EcoLexicon was built using two types of resources: manually selected domain corpora (bottom-up approach) and a collection of domain thesauri, dictionaries and lexicons (top-down ap-

[1] http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/

[2] http://ecolexicon.ugr.es/en/index.htm

[3] 22,998,357 segments for EN-DE pair; 18,764,038 for EN-FR and 5,110,262 for FR-DE (PatTR web site)

[4] International Patent Classification, http://www.wipo.int/classifications/ipc/en/

| Language | Nb of terms |
|---|---|
| FR | 640 |
| EN | 3079 |
| DE | 3713 |

Table 1: Number of terms by language in EcoLexicon

proach) (Faber et al., 2006). The multilingual corpora were built manually from reliable domain sources, taking into account multiple criteria (quantity, quality, simplicity and documentation). The domain-specific terminological resources were compared and evaluated in order to obtain a representative dataset.

## 3 Main issues

The PatTR corpus represents two main challenges: its size and its domain diversity. In fact, we can hardly estimate the amount of data for each IPC category without getting into the metadata analysis. Domain diversity can also be addressed through the metadata. However, a manual analysis is required: unless being a specialist of the IPC, one needs to manually establish a list of categories potentially corresponding to the ontology domain. Since this intervention is guided by human intuition, we need to validate the sub-corpora choice. Due to its size, the corpus is not designed to be read by a human user, so it is difficult to perform any manual check on the selected domain-specific sub-corpus. We address the validation by counting the concepts occurrences in the selected sub-corpora and checking that these occurrences belong mainly to domain-specific concepts of the ontology.

## 4 Set up

We defined a set up based on three main steps: (i) manually matching IPC categories to select the sub-corpora, (ii) counting concept occurrences in the selected sub-corpora and (iii) performing a semi-automatic validation of the concept occurrences.

### 4.1 Manual selection of IPC categories

The main challenge is to select the IPC categories that are suitable for the EcoLexicon ontology population and enrichment. As the corpus is very large, we cannot take all the data to check the concepts occurences. Therefore we started by looking up the domains defined in EcoLexicon and limited our interest to the domains enumerated in Table 2. Then we selected the IPC categories which might suit the EcoLexion ones. As one can notice, this manual correlation is subjective and not transparent, so we need an automated validation.

| IPC | EcoLexicon |
|---|---|
| C02F Treatment of water, waste water, sewage, or sludge | 3.2.5.1 Waste treatment and 3.2.5.2 Water treatment |
| B09C Disposal of solid waste; reclamation of contaminated soil | 3.2.5.4 Soil quality management |
| H01(G,M) Basic electric elements, C01G Inorganic chemistry, H02(J,M) Generation, conversion, or distribution of electric power, C25(B,C,D,F) Electrolytic or electrophoretic processes; apparatus therefor | 3.5 Energy engeneering |

Table 2: Manual IPC categories selection[5]

### 4.2 Occurrences count

We counted the occurrences of the concept labels to validate the selected sub-corpora. In fact, this approach is used to evaluate the ontology coverage regarding a domain corpus (Oostdijk et al., 2010). To do so, we lemmatized the corpus with the TreeTagger (Schmid, 1994) tool and transformed both the corpus and the concept labels to lowercase. This caused some problems, because some labels lost their domain specificity (for example, *Be@en* for *berrilium* became *be* and was found nearly in every English phrase). So we had to limit the labels to words longer than 2 characters.

We calculated the percentage of the concept occurrences in the total amount of tokens in the domain sub-corpus. For example, the English sub-corpus for the C02F category has 1,339,946 occurrences for 7,806,687 tokens, so the concept occurrences represent 17% of the tokens (the highest rate in our data collection). The least covered sub-corpus is the French H02M one with 1% of occurrences (55,803 occurrences for 4,359,434 tokens).

[5] As the category titles are too complex, we took in this table the generic IPC descriptions (i.e. *Basic electric elements* is the title of the whole H01 category)

Our hypothesis is that the sub-corpora containing more ontology concepts are more likely to be efficient for ontology population, so we will start the ontology population from the most covered sub-corpora.

The disparity in the coverage among languages observed in the Table 4 (17.16% maximum for English, 3.67% for German and 3.60% for French) can be explained by the difference in the number of EcoLexicon labels for these languages (cf Table 1). As we use a parallel corpus, we will base the suitability analysis on the occurrence percentages for English and try to find the terms translations for the other languages from the corpus.

### 4.3 Semi-automatic validation

The purpose of this step is to see which concepts appear in the corpus and to validate that their meaning in the corpus matches the one described in the ontology.

We noticed that a part of the occurrences belongs to quite general concepts that are quite close to the definition of transdisciplinary vocabulary (Tutin, 2007; Jacquey et al., 2013), such as *method, device, process* which is due to the fact that the corpus contains segments from patents. We want to be sure that the total occurrences count is not made only of these concepts. To do so, we definded a set of five recurrent concepts and their labels in the three languages (cf Table 3) in order to calculate their percentage in the total occurrences count.

| Concept | Labels |
|---|---|
| Method | method@en, mthode@fr, Methode@de |
| Process | process@en, processus@fr, Prozess@de |
| Treatment | treatment@en, traitement@fr, Verarbeitung@de, Behandlung@de |
| Device | device@en, outil@fr, Mechanismus@de |
| System | system@en, systme@fr, System@de |

Table 3: Manual concepts and labels selection

The combination of the concept occurence and the general concept percentages (cf Table 4) gives a better idea of the best sub-corpora to be used in the next steps. The highest percentage of general concepts is 19% (C25F for English), that means that almost every 5th occurrence is a general concept one. Without final results of the ontology population and enrichment, we cannot judge if this proportion is too high. The maximal percentages of the general concepts for German and French are respectively 9.14% and 1.19% of the concept occurrences.

| IPC | Lang | Occurrences % | General concepts % |
|---|---|---|---|
| C02F | **en** | **17.16** | 11.86 |
| B09C | en | 16.17 | 13.40 |
| C25C | en | 12.54 | 11.91 |
| C25D | en | 11.66 | 14.88 |
| C01G | en | 11.57 | 14.72 |
| C25B | en | 11.18 | 13.43 |
| C25F | **en** | 11.04 | **19.00** |
| H01M | en | 10.32 | 10.73 |
| H02J | en | 9.57 | 15.49 |
| H01G | en | 8.15 | 12.12 |
| H02M | en | 8.08 | 9.54 |
| B09C | **de** | **3.67** | 6.66 |
| B09C | **fr** | **3.60** | 0.99 |
| C02F | de | 3.36 | 7.29 |
| C25C | fr | 3.33 | 0.88 |
| C25C | de | 3.12 | 2.66 |
| C01G | fr | 3.10 | 0.46 |
| C25B | fr | 2.93 | 0.79 |
| H01M | fr | 2.69 | 0.55 |
| C25D | fr | 2.63 | 0.98 |
| C01G | de | 2.57 | 2.91 |
| C25F | de | 2.55 | 6.75 |
| C25D | de | 2.48 | 4.41 |
| C25B | de | 2.25 | 5.31 |
| C25F | fr | 2.18 | 1.09 |
| H01G | de | 1.94 | 2.70 |
| H01M | de | 1.86 | 4.17 |
| H01G | fr | 1.79 | 1.13 |
| H02J | **de** | 1.68 | **9.14** |
| C02F | **fr** | 1.63 | **1.19** |
| H02J | fr | 1.57 | 0.94 |
| H02M | de | 1.39 | 4.03 |
| H02M | fr | 1.28 | 0.45 |

Table 4: Concept occurrences and general concepts percentages

We also manually checked 5 random segments for 10 randomly selected terms, for example *surface water, waste, biomass, etc.*, to be sure that they preserve their terminological meaning. This quick validation helped us to confirm that the selected sub-corpora can be used for future treatments.

Regarding the meaning of the matched terms, the patent titles and abstracts preserve the terminological sense, while the claims part has more rigid style and uses some specific expressions, like *method as in claim X, product accord to one of the claim X, a process along the line of claim*, etc. In the same time, domain specific terms contained in claims can still be used as such.

## 5 Conclusion

The described set up can save time while using a large corpus for the ontology population task. The combined use of metadata and occurrences count show the best sub-corpora that we should keep for further treatment. The semi-automatic validation of occurrences is a useful step which helps to ensure that we know the data used in the project.

## Acknowlegments

The research project is supported by the CIFRE grant 2013/0744 delivered by the ANRT. We are grateful to the Lexicon research group[6] of the University of Granada for the access to the EcoLexicon ontology.

## References

[Araúz et al.2011] Pilar Araúz, Arianne Reimerink, and Pamela Faber. 2011. Environmental knowledge in EcoLexicon. In *Computational Linguistics-Applications Conference*, number 14, pages 9–16.

[Aussenac-Gilles et al.2006] Nathalie Aussenac-Gilles, Anne Condamines, and Florence Sèdes. 2006. Evolution et maintenance des ressources termino-ontologique: une question à approfondir. *Information interaction intelligence*, HS.

[Bouamor2014] Dhouha Bouamor. 2014. *Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables*. Ph.D. thesis, Université Paris Sud - Paris XI.

[Cimiano2006] Philipp Cimiano. 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Application*. Springer US.

[Delpech et al.2012] Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora : compositional translation and ranking. In *COLING*, volume 3.

[Faber et al.2005] Pamela Faber, Carlos Márquez Linares, and Miguel Vega Exposito. 2005. Framing Terminology: A Process Oriented Approach. *Meta: journal des traducteurs*, 50(4):1492–1421.

[Faber et al.2006] Pamela Faber, Silvia Montero Martínez, María Rosa Castro Prieto, José Senso Ruiz, Juan Antonio Prieto Velasco, Pilar León Arauz, Carlos Márquez Linares, and Miguel Vega Expósito. 2006. Process-oriented terminology management in the domain of Coastal Engineering. *Terminology*, 12(2):189–213.

[Faber et al.2009] Pamela Faber, Pilar Leon, and Juan Antonio Prieto. 2009. Semantic Relations, Dynamicity, And Terminological Knowledge Bases. *Current Issues in Language Studies*, 1:1–23.

[Faber2011] Pamela Faber. 2011. The dynamics of specialized knowledge representation: Simulational reconstruction or the perceptionaction interface. *Terminology*, 17(1):9–29.

[Jacquey et al.2013] Evelyne Jacquey, Agnès Tutin, Laurence Kister, Marie-paule Jacques, Sylvain Hatier, and Sandrine Ollinger. 2013. Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines. In *Terminologie et Intelligence Artificielle (TIA)*, Paris.

[Kamel et al.2013] Mouna Kamel, Nathalie Aussenac-Gilles, Davide Buscaldi, and Catherine Comparot. 2013. A semi-automatic approach for building ontologies from a collection of structured web documents. In *K-Cap'13 Proceedings of the seventh international conference on Knowledge capture*, pages 139–140.

[Oostdijk et al.2010] Nelleke Oostdijk, Suzan Verberne, and Cornelis Koster. 2010. Constructing a broad-coverage lexicon for text mining in the patent domain. In *LREC*, pages 2292–2299.

[Pinnis2014] Marcis Pinnis. 2014. Bootstrapping of a Multilingual Transliteration Dictionary for European Languages. In *Proceedings of the Sixth International Conference Baltic HLT*.

[Schmid1994] Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.

[Tutin2007] Agnès Tutin. 2007. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, XII:5–14.

[Wäschle and Riezler2012] Katharina Wäschle and Stefan Riezler. 2012. Structural and Topical Dimensions in Multi-Task Patent Translation. In *The 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 818–828, Avignon, France.

[6] http://lexicon.ugr.es/