Corpus-driven Terminology Work for Describing Basque Academic Terminology: the Weaving Terminology Networks programme (TSE programme)

Igone Zabala, Izaskun Aldezabal, Maria Jesús Aranzabe, Jose Maria Arriola.
Itziar Gonzalez-Dios, Mikel Lersundi
Ixa NLP group, Basque Language and Communication Department
University of the Basque Country (UPV/EHU)

References


Background

- Languages undergoing normalization processes require terminology planning. Nevertheless, natural development and self-regulation inside specialized discourse communities also contribute to the development and normalization of terminology (Cabré 2003).
- In well-developed languages, terms present both formal and conceptual variation, due to cognitive, dialectal, functional, discursive, and interpersonal causes (Freixa 2003).
- Sociolinguistically unstable minority languages present a greater frequency of free variation than well-developed languages, but they also show pragmatically motivated variation due to discursive, functional, or cognitive causes. Terminology description based on functional criteria is required in order to detect and distinguish both kinds of variation (Elordui & Zabala 2005).
- Functionally motivated variation should be promoted by normalizing initiatives, and unsystematic variation should be harmonized, in order to enrich pragmatically and cognitively motivated variation (Elordui & Zabala 2005, Zabala 2018).

Weaving Terminology Networks programme (University of the Basque Country UPV/EHU, since 2008)

Basque Language and Communication Department

Motivation:
- Basque intensively used as an instruction and communication language in university courses in all kinds of disciplines, doctoral theses, master’s theses and a large number of end-of-degree works.
- Terminology used in academic settings should be described and analysed since the expert’s real usage in terminology planning may hinder the natural development of academic registers of Basque.
- There is a fluid and well-established communication between teachers and students in Basque, but lack of fluid communication networks in Basque among experts/teachers (Zabala et al. 2014).

Goals:
- To compensate for the lack of fluid communication in Basque among experts/teachers.
- Monitoring texts and terminology used in academic communication in Basque.

Methodology:
- Work-unit: teaching subject = communication unit
- Active description: the authors of the texts describe the terminology they use in their own texts with the help and guidance of linguists involved in the programme.
- Corpus-driven terminology work
- A bottom-up approach to concepts

Process/ Working flow to create dictionaries by experts
1. Adding texts to Garacterm corpus
2. Extracting term proposals automatically from the added texts
3. Validating term proposals
4. Creating Basque term list: correcting orthography, adding variants...
5. Adding term equivalents in other languages
6. Uploading multilingual term list to TZOS
7. Grouping terms by semantic classes
8. Adding definitions
9. Profiling the teaching subject
10. Adding semantic classes and definitions to TZOS

Publically available resources:
- Garaterm corpus: linguistically processed 18 023 178 text-words, 250 authors, 25 subject domains.
- TZOS (Online System for Terminology Service): 32 438 term-entries; 60 subject domains and 321 sub-domains; 151 authors

Applications:
- Evaluation of the implementation of terminology standardized by normalizing institutions: Terminologia Batzordea (Terminology Commission Euskaltzaindia (Basque Language Academy)
- Detection of terms and variants not codified in databases and dictionaries

Improvement of experts’ linguistic competence:
- Weaving Terminology Networks (TSE) is a training program that contributes to the dissemination of terminology standardized by normalizing institution.

Conclusion and Future Works

- Thanks to the TSE programme a considerable amount of terms not available in Euskaltzaindia (the Basque Public Terminology Bank) are being described.
- Academic subject, an interesting work unit: the terminology used in different subjects of the same domain is quite different.
- An university subject can be classified inside a domain, but several terms traditionally classified in other domains are also used.
- The semantic category assigned to terms by experts depends more on the viewpoint of the subject than in the concept system of the traditional domain.
- Future work: comparing bottom-up semantic analysis with top-down knowledge representations
- Harmonization of variants via aclatory notes.

Comparative results

Figure 1: Percentages of terms described with TSE methodology from teaching materials of the two university subjects:
- Terminiologia Histolohikoa eta Zaleurraren Histolohiko Teknikak eta Cellu Gaitasuna.
- Optika: Optikea.
- Experrimentazioa Kimiko Fisikoa: "Experimentation in Physical Chemistry".
- Análisi Matemàtikoa eta Kalkulu: Mathematik Analysis and Calculus.

Figure 2: Semantic profiles of described terminology: Biokimikoa/Basic Biochemistry vs Omurtan Metabolotegia Biokimikoa/Basic Biochemistry Methodology.

Figure 3: Distribution among different domains of the terms described in Biochemistry (BBM) and Basic Biochemistry Methodology (BBM), which are codified in Euskaltzaindia.

References