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

Igone Zabala, Izaskun Aldezabal, María 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)
igone.zabala@ehu.eus, izaskun.aldezabal@ehu.eus, maxux.aranzabe@ehu.eus,
josemaria.arriola@ehu.eus, itziar.gonzalezd@ehu.eus, mikel.lersundi@ehu.eus

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 interlinguistic causes (Freixa 2003).
- **Sociolinguistically unstable minoritized 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 in 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).

Terminologia Sareak Ehunduz

Weaving Terminology Networks programme

(University of the Basque Country UPV/EHU, since 2008)

Basque Language and Communication Department

Motivation:

- o **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.
- o 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.
- o 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:

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

Methodology:

o Work-unit: teaching subject = communication unit

o 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.

o Corpus-driven terminology work

o A bottom-up approach to concepts

Process/ Working flow to create dictionaries by experts

1. Adding texts to Garaterm 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

Applied results

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.

Comparative results

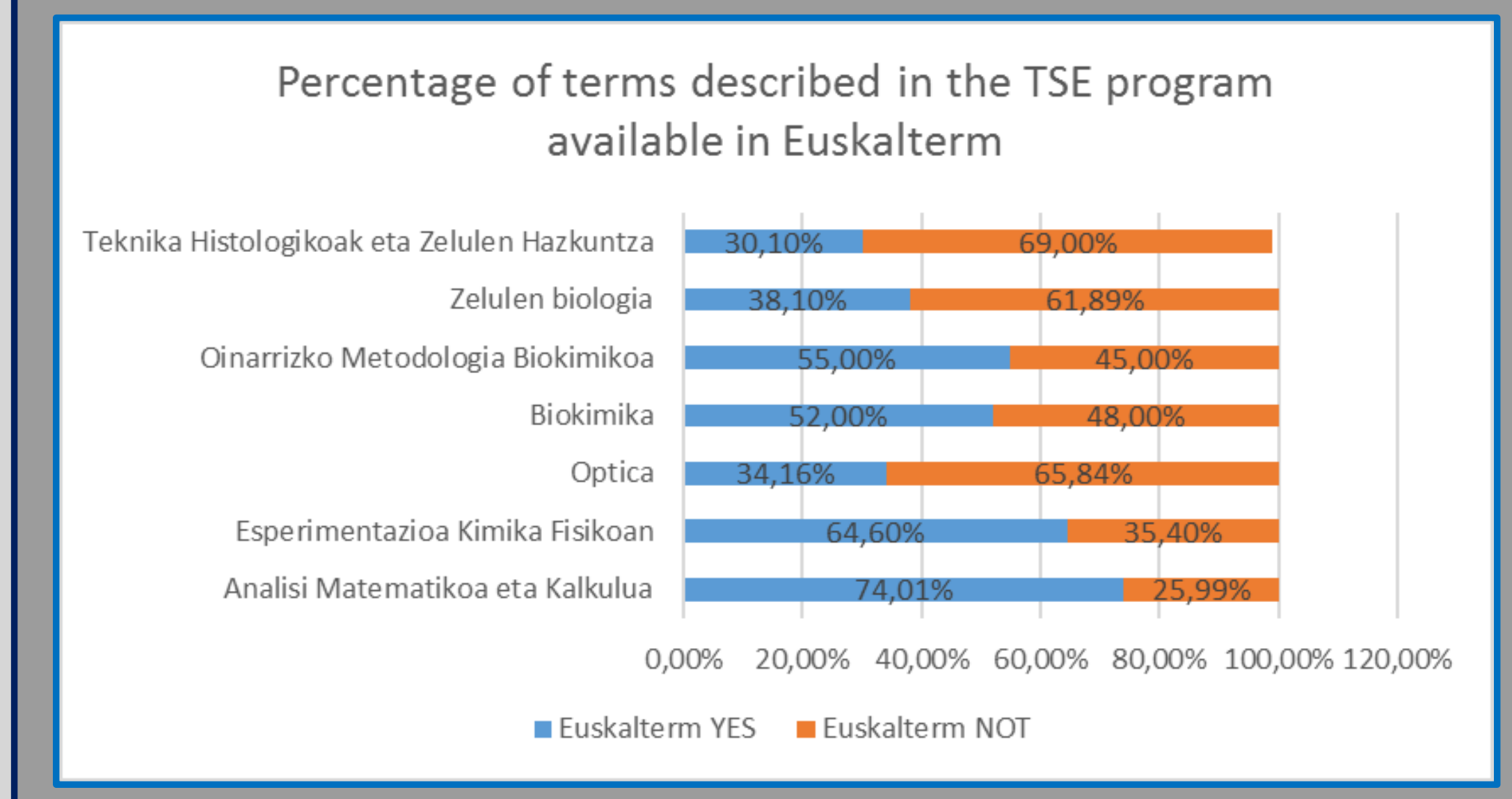
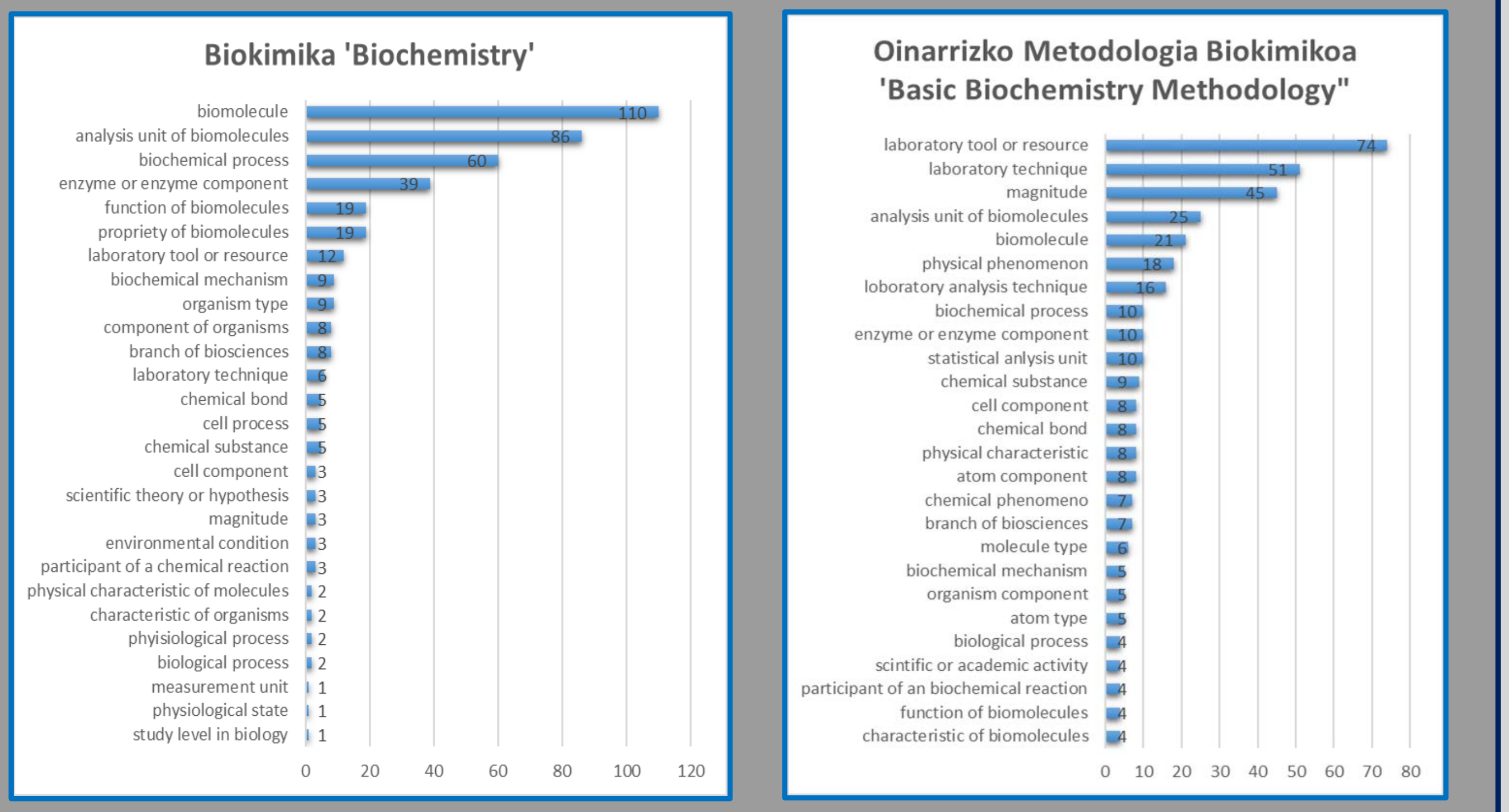


Figure 1: Percentages of terms described with TSE methodology from teaching materials of this university subjects:
- *Teknika Histologikoak eta Zelulen Hazkuntza* 'Histologic Techniques and Cell Culture'
- *Zelulen Biologia* 'Cell Biology'
- *Oinarrizko Metodologia Biokimikoa* 'Basic Biochemistry Methodology'
- *Optika* 'Optics'
- Esperimentazioa Kimika Fisikoan 'Experimentation in Physical Chemistry'
- *Analisi Matematikoa eta Kalkulua* 'Mathematic Analysis and Calculus'

Figure 2: Semantic profiles of described terminology: *Biokimika* 'Biochemistry' vs *Oinarrizko Metodologia Biokimikoa* 'Basic Biochemistry Methodology'



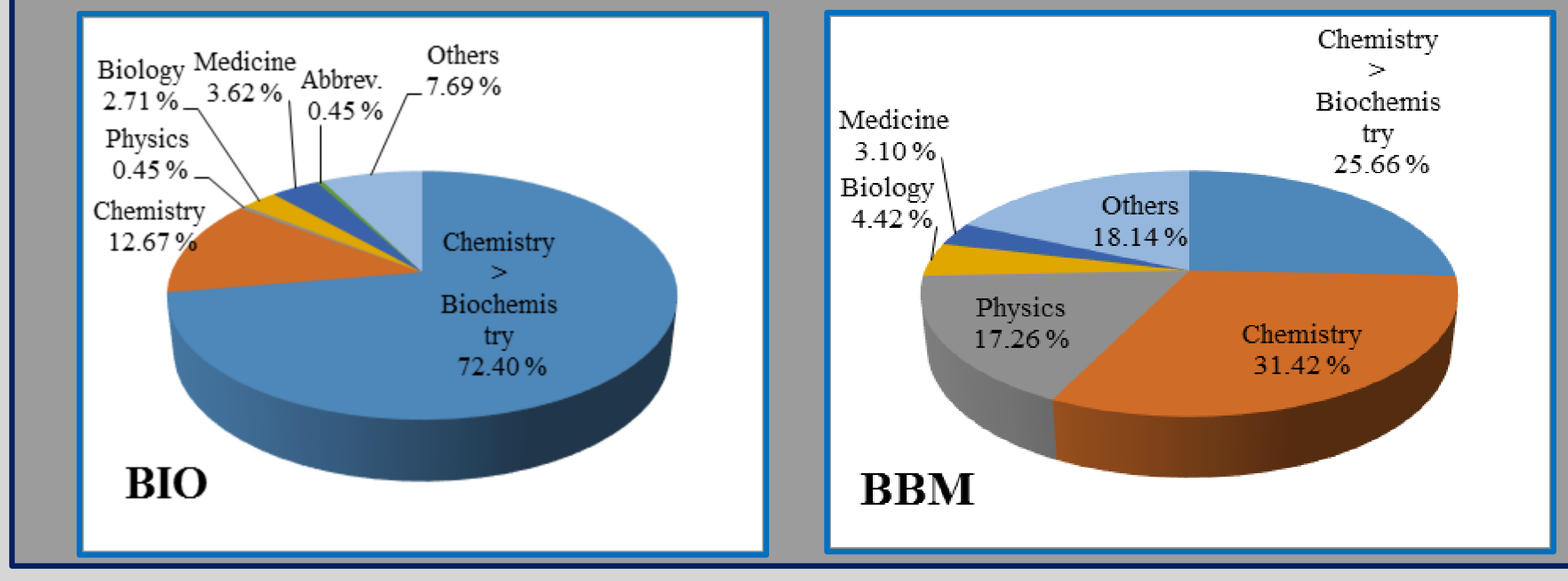


Figure 3: Distribution among different domains of the terms described in Biochemistry (BIO) and Basic Biochemistry Methodology (BBM), which are codified in *Euskalterm*.

Conclusion and Future Works

- Thanks to the TSE programme a considerable amount of terms not available in *Euskalterm* (the Basque Public Terminology Bank) are being described.
- Academic subject, an interesting work unit:
  - o Terminology used in different subjects of the same domain is quite different.
  - o An university subject can be classified inside a domain, but several terms traditionally classified in other domains are also used.
  - o 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:
  - o Comparing botton-up semantic anlysis with top-down knowledge representations.
  - o Harmonization of variants via aclaratory notes.

References

- Cabré. Mª T. 2003. Terminología y normalización lingüística. In Xabier Alberdi, Iñaki Ugarteburu & Pello Salaburu (eds.), *Espezialitate hizkerak eta terminologia jardunaldiak*, 11-25. Bibao: Publishing Service of the UPV/EHU.
- Freixa, J. 2003. *La variación terminológica: Anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'area de medi ambient*. Barcelona: Universitat Pompeu Fabra dissertation.
- Elordui, A. & Zabala. I. 2005. Terminological variation in Basque: Analysis of texts of different degrees of specialization. *SKY Journal of Linguistics* 18. 71-91.
- Zabala, I., San Martin, I. & Lersundi, M. 2014. Linguistic and sociolinguistic factors that influence the detection, implantation and circulation of natural terminology in academic uses of Basque. In Pascaline Dury, Jose Carlos de Hoyos, Julie Makri-Morel, François Maniez, Vincent Renner & Maria Belén Villar Díaz (eds.), *La néologie en langue de spécialité. Neology in specialized languages. La neología en lengua de especialidad*, 141-164. Lyon: Publications du CRTT.
- Zabala, I. 2018. Euskararen lexiko espezializatuaren garapenaz eta harmonizazioaz. In Gidor Bilbao, Pruden Gartzia eta Mari Karmen Garmendia (eds.), *Bai, jauna, bai: fisika euskaraz! Jose Ramon Etxebarria irakaslearen omenez, 349-358*. Bilbo: UEU.