8-27-2024

Computational Approaches to Bilingualism in Spanish Education: a Database for Natural Science Students

Federico Ortega Riba

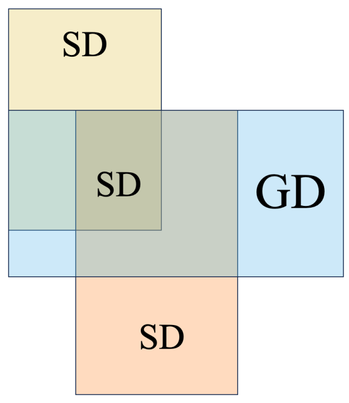
Federico Ortega Riba - Fulbright Grantee 2024-25
federicoortegariba@gmail.com

Computational Approaches to Bilingualism in Spanish Education: a Database for Natural Science Students

1. Introduction

One major issue in creating dictionaries, especially for young students is **balancing general and specialized language**. Computational tools can provide the most frequent single and multi-word terms in a specialty field to help students learn. However, in primary education, the most frequent words might not appear specialized at first glance.

Within **specialized discourse**, sub-discourses can develop based on the knowledge of the student. For example, the language level between two ten-year-old students differs from that between a ten-year-old and a seven-year-old. Cabré (1993) argues there is an **intersection between general and specialized languages** due to these pragmatic dependencies.

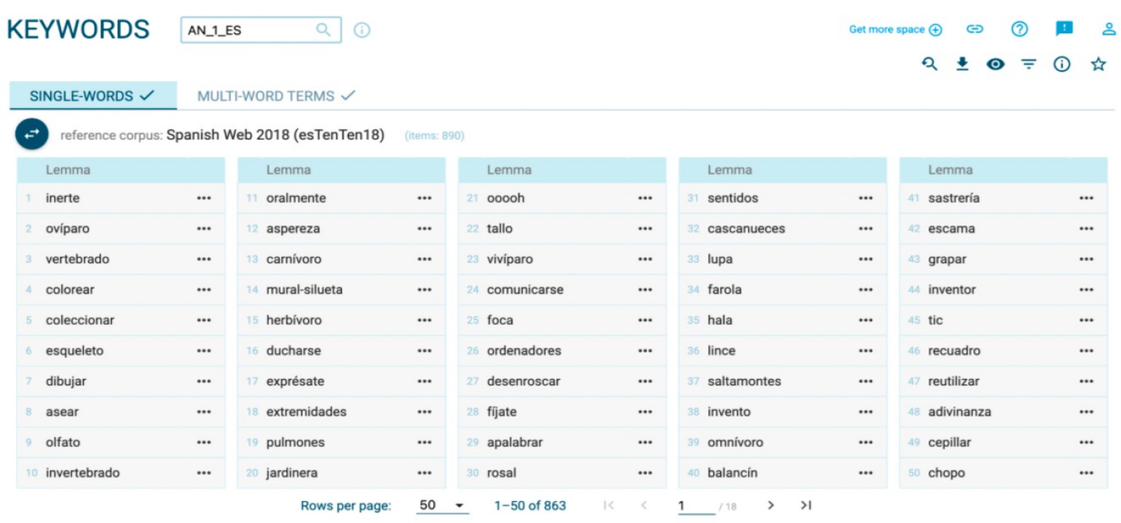


Specialized discourse (SD) and General discourse (GD). Adapted from Gómez (2005, 46).

2. Methods

Keyword function in Sketch Engine & logo

- A comparable corpus has been compiled for each grade level of the Spanish Anaya (2019) edition for the subject of **Natural Sciences** and its English version.
- Additionally, each compiled academic unit follows this labeling format: *topic_publisher_language_grade_subdomain_extension*.
- After extracting the units in *.txt* format for each grade, they were uploaded to the **Sketch Engine (SK)** tool. Each corpus follows this labeling format: publisher_grade_language, and both single and multi-word terms in Spanish and English have been extracted.



Single words	Frequency
viviparous	6
oviparous	6
omnivore	6
vertebrate	6
invertebrate	6
amphibian	6
reptile	6
herbivore	5
carnivore	5
pollute	5
intestine	5

- The number of keywords has been limited to **120** for first grade and **200** for the other levels, in both the English and Spanish corpora.
- The resulting symmetrical terms in both collections have been compared in Excel to extract those with a **frequency range from 3 to 6** in all corpora.

***Above**: Keywords function in SK*
***Left**: Results of keywords using the COUNTIF function in Excel*

There is **more than a 50% correlation** between the frequency of terms in the Spanish corpus and their equivalents in the English corpus: specifically, **60.31%** for single terms and **69.33%** for multi-word terms.

3. Results

A public domain on the **WordPress.com** platform was used to create the database. The design of the entries is organized as follows:

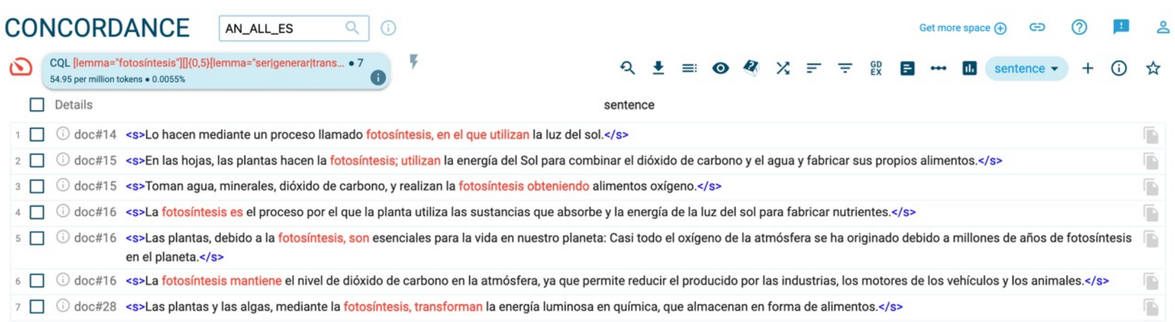
- Definition and example of use
- Equivalent
- Collocations and example of use
- Image
- References

RegEx

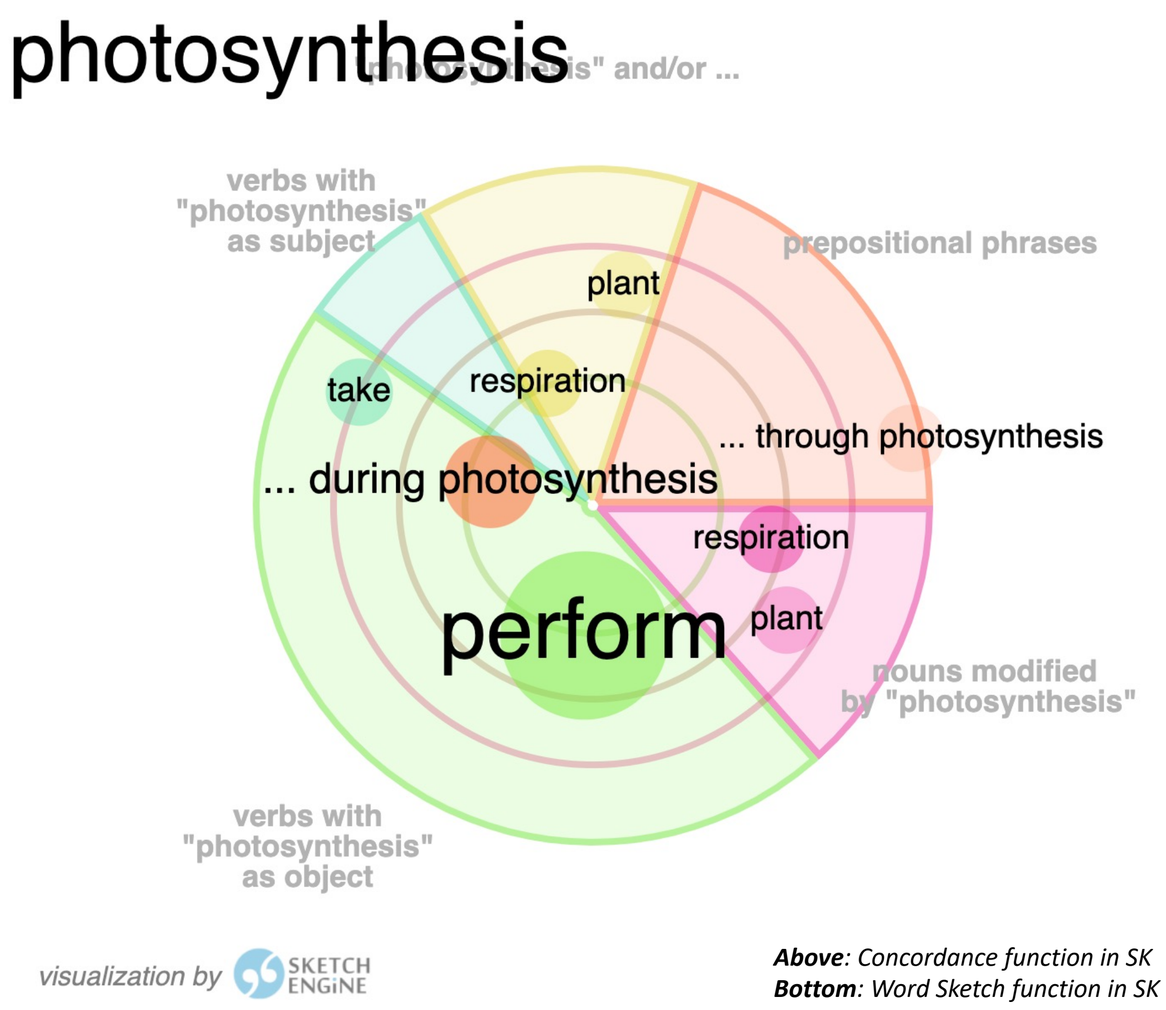
This illustration was made by M0tty

Let us study how an example of entry for the term ***fotosíntesis*** and its English equivalent is outlined.

1. A search is conducted in **CQL** using the following RegEx: [lemma="fotosíntesis"][]{0,5}[lemma="ser|generar|transformar|obtener|utilizar|mantener"] within <s/>.



2. **Collocations** are extracted using the Word Sketch function.
3. **Definitions** are extracted from the CQL search.
4. The **image** has a Creative Commons license and it is extracted from the internet.



***Above**: Concordance function in SK*
***Bottom**: Word Sketch function in SK*

4. Conclusion

- There is currently a need for the creation of **open educational resources**.
- The database is proposed as a reference work that will address the **limitations of bilingual education** in Spanish Primary Education.
- This research is seen as a **precursor** to a Master's dissertation or even a doctoral dissertation.

5. References

SCAN ME!