# Corpora used for the *Representativeness heuristics* test

May 2024

Italian corpora:

- ITWAC: `https://bellatrix.sslmit.unibo.it/noske/public/#wordlist?corpname=itwac_full&tab=basic&wlattr=lc&include_nonwords=1&itemsPerPage=50&showresults=1&cols=%5B%22frq%22%5D&search_query=s`
- Paisà: `https://www.corpusitaliano.it/it/access/simple_interface.php`

English corpora:

- UKWAC: `https://bellatrix.sslmit.unibo.it/noske/public/#wordlist?corpname=ukwac_full&tab=basic&wlattr=lc&include_nonwords=1&itemsPerPage=50&showresults=1&cols=%5B%22frq%22%5D&search_query=no`
- NOW: `https://www.english-corpora.org/now/` requires login.

Spanish corpora:

- NOW: `https://www.corpusdelespanol.org/now/` requires login.
- Web: `https://www.corpusdelespanol.org/web-dial/` requires login.

| Corpus name | Language | Training data | Yes occurrences | No occurrences |
|---|---|---|---|---|
| ITWAC | Italian | 1bln | 300.668 | 538.060 |
| Paisà | Italian | 250mln | 15.982 | 24.166 |
| UKWAC | English | 1bln | 263.241 | 2.705.457 |
| NOW | English | 18.8bln | 1.853.453 | 24.489.421 |
| NOW | Spanish | 7.3bln | 2.403.314 | 50.591.620 |
| Web | Spanish | 2bln | 1.147.884 | 22.061.031 |

Table 1: Corpora occurrences of "yes" and "no"