**FINAL SCIENTIFIC REPORT**

# ‘Getting oriented in the labyrinth: A cross–linguistic topographic map of the human lexicon’

**Davide Crepaldi**

MoMo Lab
Department of Psychology, University of Milano-Bicocca
20126, Milan, Italy
`davide.crepaldi1@unimib.it`
`www.davidecrepaldi.net`

**Hosting Institution:** Department of Psychology, Universiteit Gent, Belgium
**Visiting period:** 8 October 2012 to 19 October 2012

## 1 Goal of the project

Word features and psycholinguistic properties (e.g., length, frequency, number of orthographic neighbours, morphological structure) have very different distributions across languages, which might well be relevant for mental processing because all these variables were shown to impact heavily on word processing time and quality (Baayen et al., 2006). The broad goal of this project is to develop *lexical topographic maps* for different languages by conceiving these variables as defining an N–dimensional space where words (represented as individual points) distribute. These maps would surely be topographically different in different languages, which might help explaining cross–linguistic inconsistencies. More in the long run, this approach might establish as a new research method itself: cognitive systems are organized optimally to deal with their everyday–life input, and so a detailed and formalized description of this input might help revealing aspects of their structure.

## 2 Purpose of the visit

My collaborators on this grant, Marc Brysbaert and Emmanuel Keuleers, have developed over the last years frequency databases for several different languages (e.g., English, Spanish, German, Dutch) based on movie subtitles (*http://crr.ugent.be/programs-data/subtitle-frequencies*). These new databases feature two big moves forward, i.e., (i) being based on material that is more representative of everyday–life language, they explain human performance better than previous databases (Brysbaert and New, 2009) and (ii) they have been developed through a standard workflow, which guarantees comparability between languages and makes it likely that more languages will join the club in the near future. Because reliable databases are of vital importance for the goal of the project, the purpose of my visit was to (i) get accustomed with the existing databases; (ii) familiarize with the workflow that

Figure 1: Frequency distribution in SUBTLEX–IT (blue) and CoLFIS (pink).

generated those databases; (iii) explore the possibility of developing a frequency database for Italian; (iv) start to develop the statistical–mathematical tools to describe the lexical space as an N–dimensional geomtric space; (v) discuss possible dissemination strategies that would put the outcome of this work in the grasp of other researcher, professionals, and the public.

# 3 Description of the work carried out during the visit

Prior to the beginning of the visit, I downloaded the existing databases that were developed by Dr. Keuleers and Prof. Brysbaert, read the paper where they were illustrated (Brysbaert and New, 2009; Cai and Brysbaert, 2010; Keuleers et al., 2010a, e.g.,), and started to develop processing routines, based on the *R* software (R Development Core Team, 2011), to manipulate them. The first task that I took up directly during my visit to Gent was then to explore the possibility of developing a database for Italian; we thought that, in addition to addressing directly goal (iii), this was also the best way for me to familiarize with the workflow. With the additional (and very substantial) help of Paweł Mandera, we (i) surfed the internet in search of Italian subtitle text files; (ii) applied an algorithm to guarantee that no pairs of files referred to the same movie; (iii) tokenized each file, i.e., segmented the text corpuses into lists of individual words; (iv) counted how many times each word occurred in the whole corpus (frequency count); (v) counted how many different movies (i.e., subtitle files) each word appeared in (contextual diversity count). These processing steps capitalized on tools that were already available in the lab from previous work on other languages, with substantial time savings. However, subtitle files always contain transcription errors, occasionally in relevant proportions. It was thus necessary to apply my knowledge of the language as a native Italian speaker to (i) spot these errors in the word lists; (ii) figure out whether they were occasional or recurrent errors; (iii) if the latter, conceive heuristics to correct them automatically within the tokenization script. This process was quite time consuming because it required several cycles of generating a new frequency list and testing it, which of course could not be done but through hand checking each item individually.

We were also interested in attaching part–of–speech tags (e.g., noun, verb, adjective) to each entry in the frequency list; because there is abundant evidence that words of different grammatical classes give rise to different behavioural phenomena (Mahon et al., 2007), we thought it was worth to collect this kind of information so as to explore the possibility of developing different lexical maps for different parts of speech in different languages. We then explored the several tools that are available on the internet to automatize part–of–speech tagging, to see whether some of them was able to handle Italian. We found two such tools, i.e., *TreeTagger* and *FreeLing*, tried them on our corpora, and compared their performance. Both these tools also features lemmatization routines, i.e., scripts that assign each word form to its lemma, or base form (e.g., *cats* is a form of the lemma *cat*, *bought* of the lemma *buy*). We thus obtained also this type of information for each entry in the list.

# 4 Description of the main results obtained

The result of the work described above is a list of Italian words with PoS tags, lemmas, frequency and contextual diversity measures. PoS tags are based on *TreeTagger* because this

Figure 2: Proportion of $X$ CoLFIS highest–frequency words that are also among the $X$ SUBTLEX–IT highest–frequency words, as $X$ increases.

tool turned out to yield more accurate results than *FreeLing*, as attested through hand tagging on 200–words samples (95% confidence interval estimates are $.917 - .952$ vs. $.863 - .908$). Based on these four basic variables, we also developed additional metrics for each entry, i.e., the frequency of the dominant lemma, the frequency of the dominant PoS tag, the relative incidence of the dominant lemma, and the relative incidence of the dominant PoS tag. Potentially, it would be easy to get measures for lexical–syntactic entropy, i.e., how much the dominant tag is dominant compared to the alternative tags, an index which might be interesting to explore in the future (Moscoso del Prado Martín et al., 2004). Of course, the current version of the list is not the final one, as the tokenization, lemmatization and counting routines are still being improved. However, I am able to offer some quantitative comparison between the current film–based database (dubbed SUBTLEX–IT in what follows) and the best available frequency database in Italian, the *CoLFIS* (Bertinetto et al., 2005). A qualitative test will only be possible by comparing the proportion of variance explained by frequency in the two databases against real behavioural data, whose collection is being planned, but did not started yet (see Section 5).

SUBTLEX–IT is based on a 129433373–words corpus, while CoLFIS figures were computed on the basis of a sample of 3982442. This makes SUBTLEX–IT more sensible in the low–frequency range (see Figure 1) or, more specifically, drives SUBTLEX–IT to (i) give more precise estimations of low frequency values; (ii) differentiate words more subtly in that frequency range; (iii) more in general, yield lower frequency values than CoLFIS. These features of SUBTLEX–IT might be very relevant given that recent findings have shown that the frequency effect arises more strongly at this frequency range (Keuleers et al., 2010b).

SUBTLEX–IT feature 550539 single entries, while CoLFIS features 65539. These absolute figures do not tell much: both lists include punctuation marks, which clearly are not words[1], and we are still working on the processing routines to make sure that SUBTLEX–IT isn't generating false words[2]. However, a comparison between them suggests that the significantly larger (and arguably more representative) corpus on which SUBTLEX–IT is based might lead to a better coverage of all real existing Italian words and to a more precise differentiation between words.

Finally, some notes on how much SUBTLEX–IT and CoLFIS figures correlate. Figure 2 indicates the proportion of $X$ CoLFIS highest–frequency words that are also among the $X$ SUBTLEX–IT highest–frequency words, as $X$ increases. Half of the 100 highest–frequency words in Italian according to CoLFIS are also among the 100 highest–frequency words in Italian according to SUBTLEX–IT; this proportion jumps to around 66% when the 500 highest–frequency words are considered, and then declines progressively to around 56% when the whole CoLFIS database is considered. The overall correlation between the frequency figures (occurrences per million) of the words that are included in both databases is .57. On the one hand, these data indicate that SUBTLEX–IT is providing new information as compared to CoLFIS (correlation isn't excessively high); on the other hand, however, these consistency measures also suggest that the new figures are likely to be reliable, particularly when one considers the highest–frequency words (correlation isn't excessively low).

[1] We plan to clean them out from the final version of SUBTLEX–IT.

[2] The total number of entries is suspiciously high also compared to the other SUTBLEX databases, e.g., the Dutch list includes 134723 entries and the American English list includes 74286.

## 5 Future collaborations with host institution

There are two further steps in the project that are already in our plans. First, we need to run a validation study to assess how these new frequency figures perform in accounting for variance in real human data, particularly in comparison with existing databases (such as CoLFIS). Given the people in Keuleers and Brysbaert's lab (notably, Micheal Stevens) have developed software that might be used to gather behavioural data on a large sample of words by a large sample of readers, we are planning to do this by running a mega–study on lexical decision times with students at the University of Milano Bicocca. Second, we still need to explore statistical methods to define the topographic space, once the database will be finalized.

## 6 Projected publications

It is difficult to provide solid predictions about eventual publications at this stage, but, also on the basis on the dissemination plan followed by Dr. Keuleers and Prof. Brysbaert with the other SUBTLEX databases, it is likely that the project will result in two papers, one that will describe the new frequency database for Italian and one that will report on topographic maps as tools to understand more deeply the experimental data available on how humans identify printed words.

## References

Baayen, R. H., Feldman, L. B., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53:496–512.

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., and Thornton, A. M. (2005). Corpus e lessico di frequenza dell'italiano scritto (CoLFIS).

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for Aerican English. *Behavior Research Methods*, 41:977–990.

Cai, Q. and Brysbaert, M. (2010). SUBTLEX–CH: Chinese word and character frequencies based on film subtitles. *Plos ONE*, 5:e10729.

Keuleers, E., Brysbaert, M., and New, B. (2010a). SUBTLEX–NL: A new frequency measure for dutch words based on film subtitles. *Behavior Research Methods*, 42:643–650.

Keuleers, E., Diependaele, K., and Brysbaert, M. (2010b). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14000 dutch mono– and disyllabic words and nonwords. *Frontiers in Psychology*, 1:174.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., and Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, . 33(3):503–535.

Moscoso del Prado Martín, F., Kostic, A., and Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94:1–18.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien, Austria.