Automating psycholinguistic statistics computation: Procura-Palavras

João F. Machado[1], *José João Almeida*[1], *Alberto Simões*[1], *Ana Soares*[2]

[1]Departamento de Informática, Universidade do Minho, Portugal
[2]Escola de Psicologia, Universidade do Minho, Portugal
joaoffm@gmail.com, jj@di.uminho.pt, ambs@di.uminho.pt, asoares@psi.uminho.pt

Abstract

This article describes psycholinguistic lexical databases available in various languages, including English, Spanish and Portuguese. These lexical databases are important for researchers in Psycholinguistics and other related areas, providing a pool of experimental materials and allowing for an efficient process of selection of these experimental materials.

The process of gathering statistics is slow, resulting in a small pool of materials in the short-term. The need to find an alternative method to gather limited or yet unavailable statistics for a specific language led us to consider gathering statistics from other languages and to compute their triangulation. Our aim was to automatize the computation of statistics such as Familiarity, Imageability, Age of Acquisition and Written Word Frequency for that specific language.

We will describe the process of preparing this data and triangulating and comparing statistics for some languages in an attempt of finding a relationship between them. The results were analysed considering correlations between each statistic in each pair of languages and by computing the mean of absolute differences between each language's values.

Index Terms: psycholinguistic, lexical databases, psychology, linguistics

1. Introduction

Psycholinguistics is an interdisciplinary area related to various fields, such as psychology, cognitive science and linguistics. It is the study of the process by which the human mind understands language.

For those planning studies on linguistic processing, an important requirement is the undeniable need for lexical databases. These databases are the foundation of most psycholinguistic studies and they can have a great impact in the quantity and quality of those studies, providing a large pool of experimental material and allowing for a strict selection of that material.

This paper's main objective is to analyse available resources, not only for the European Portuguese language, but also for various others languages. This analysis will aid in the development of a tool that will support investigation and will be of crucial importance mainly in the area of Psycholinguistics, but also for other areas such as Cognitive Psychology, Neurosciences or Artificial Intelligence.

Despite being primarily a survey, it also contains experiments aiming to investigate the possibility of automating the computation of Portuguese words' psycholinguistic statistics based on other languages' statistics and determining the reliability of those computations. This research main goal is to report the study of an alternate method of obtaining statistics with some degree of confidence, that will allow a potentially larger pool of estimated statistics available to researchers, when none or few statistics are available.

2. Lexical databases

A key component for conducting a thorough Psycholinguistic investigation involving linguistic stimuli is the availability of comprehensive software applications that enable researchers to compute relevant psycholinguistic statistics based on lexical resources.

2.1. English resources

The English language has an application and bundled database called *N-Watch* [1], a simple tool for obtaining a broad range of lexical statistics. It provides measures of word frequency, orthographic similarity, orthographic and phonological structure, age of acquisition and imageability. The default vocabulary of 30 605 words was obtained from the *CELEX ECT* [2], a corpus of 17.9 million words.

2.2. Spanish resources

An adaptation of the *N-Watch* application was prepared for the Spanish Language, named *BuscaPalabras* (B_PAL) [3], which includes measures of word, syllable, bigram and biphone frequencies, orthographic similarity, orthographic and phonological structure, concreteness, familiarity, imageability, valence, arousal and age of acquisition. It features some important differences from the original English program, such as support for the Spanish orthographic system, statistics related to syllable measures, and lastly it enables user-defined statistics. The default vocabulary of 31 491 words was obtained from *LEXESP* [4], a corpus of approximately 5 million words.

2.3. Portuguese resources

In 2003 a European Portuguese lexical database called PORLEX [5] was made available. It provides a series of psycholinguistic statistics, and although it contains information for a total of 29 238 words, it has several limitations. The lexical frequency value is only available for 5% of those words, and it lacks semantic information and subjective psycholinguistic statistics that recent research [6] has proven to be of great importance.

Another research focuses on rated age of acquisition norms and their relation with other psycholinguistic statistics [6]. It contains a database of 834 nouns that includes age of acquisition information but also imageability, familiarity, written word frequency, concreteness, number of syllables and number of words.

In 2000, the project *Multifunctional computational lexicon*

of contemporary Portuguese [7] was concluded, with a default vocabulary of 26 443 words and 140 315 lemmas. It added much needed frequency values based on a large and diverse corpus, called CORLEX [8]. This corpus includes 16 210 438 words (95% of written corpora and 5% from oral corpora).

2.4. The project *Procura palavras*

As opposed to other languages, the current databases available for European Portuguese (eg. PORLEX, CORLEX) are outdated, limited or small, especially regarding lexical frequency and subjective psycholinguistic statistics.

In light of these conditions, we are initiating a project called *Procura Palavras* (*P-PAL*)[1], whose main goal is the development of a multi-platform software application that enables researchers to easily and simultaneously compute a broad range of objective and subjective linguistic and psycholinguistic statistics.

3. Triangulating psycholinguistic statistics

From the analysed resources we can argue that the current European Portuguese databases are unsatisfactory regarding psycholinguistic statistics. Although a gathering of psycholinguistic statistics is currently in motion, this process is extensive and time-consuming. In order to prepare a base set of statistics in less time another solution came to light: computing Portuguese statistics using other languages statistics.

For this process to be attainable with some degree of confidence, some crucial steps are required, including:

- normalizing the statistics of each language so they can be comparable (as different languages use different scale ranges),
- importing statistics to a single database for proper querying,
- connecting statistics trough translations from English (EN) to Portuguese (PT), English to Spanish (ES), Spanish to Portuguese and Spanish to English,
- filtering out possible erroneous or low confidence translations,
- determining correlations[2] and mean[3] of absolute differences between each language values,
- computing initial values for statistics with valid correlations.

For this article we selected a few psycholinguistic statistics: Familiarity (FAM), Imageability (IMG), Age of Acquisition (AoA) and Written Word Frequency (WWF):

- *Familiarity* is a statistic measured by asking readers to rate theirs familiarity with a word. A common query is to ask readers to rate how familiar they are with a specific word, measuring this familiarity on a scale of 1 (very unfamiliar) to 7 (very familiar).
- *Imageability* is a statistic measured by asking readers to rate how well they can form an image of that word in their heads, and rating it in a scale of 1 (impossible) to 7 (very easy).
- *Age of Acquisition* is a statistic measured by asking readers to estimate the age at which they think they have learned the real meaning for a word and, for example, estimating it on a 7-point scale (0-2, 3-4, 5-6, 7-8, 9-10, 11-12, and 13 or more years).
- *Written Word Frequency* is the frequency with which a word appears in a written corpus.

[1] `http://natura.di.uminho.pt/p-pal/`

[2] the degree to which two or more attributes or measurements on the same group of elements show a tendency to vary together

[3] something having a position, quality, or condition midway between extremes; a medium

3.1. Available statistics

This section shows the available statistics for each language, that are relevant for the research, including the English language with statistics from the *N-Watch* application [1], the Spanish language with statistics from the *BuscaPalabras* application [3], and the Portuguese language with statistics from the paper *Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables* [6]. All this information is summarized in Table 1.

Table 1: Language (Lang), Statistic (Stat), Number of Words (N), Range from (F), and Range to (T) for: Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Written Word Frequency (WWF). AOA range in years (Y).

Lang	Stat	N	F	T
EN	AOA	3 136	100 ($Y \leq 2$)	700 ($Y \geq 13$)
EN	FAM	4 944	100	700
EN	IMG	4 944	100	700
EN	WWF	30 591	0	1.000.000
ES	AOA	139	1 ($Y \leq 1$)	11 ($Y \geq 11$)
ES	FAM	6 223	1	7
ES	IMG	6 096	1	7
ES	WWF	31 491	0	1 000 000
PT	AOA	834	1 ($Y \leq 2$)	7 ($Y \geq 13$), 8
PT	FAM	808	1	5
PT	IMG	249	1	7
PT	WWF	790	0	15 354 243

To make this table easier to understand consider the following explanation: there are 834 words available from the Portuguese Database that have values for at least one of the statistics considered. AoA rates are available for all 834 words, range from 1 (2 years old or less) to 7 (13 years old or more) and include an eight extra point (meaning learned in adulthood). FAM rates are available for 808 words and range from 1 (highly familiar) to 5 (very unfamiliar). IMG rates are provided for 249 words ranging from 1 (smaller imageability) to 7 (greater imageability), and 790 words have WWF measures per 15 million.

3.2. Triangulation Procedure

In order to process these statistics there was the need to import them to a common database, allowing for a proper and simpler querying and providing a meticulous analysis. This was done by developing a Perl script to parse each language database and import their data to a MySQL database.

An additional task consisted in normalizing statistics, as they have different ranges for each language. Table 2 presents the normalization formula used for each language and its resulting normalized range. Without this normalization, values would not be comparable and no analysis could have been performed.

Given the difficulty to compare WWF among languages the Logarithm of Written Word Frequency (LOG-WWF) was computed. LOG-WWF ranged from -2.81 to 13.81. To make it

Table 2: Language (Lang), Normalization result (N) and Normalization formula (F) for ES and PT databases: Familiarity (FAM), Age of Acquisition (AOA), Imageability (IMG) and Written Word Frequency (WWF)

Lang	Statistic	F	N
ES	FAM	$R \times 100$	100–700
ES	IMG	$R \times 100$	100–700
ES	AOA	$((R * \frac{1}{2}) + \frac{1}{2}) \times 100$	100–700
PT	FAM	$((5 - R) \times \frac{3}{2} + 1) \times 100$	100–700
PT	IMG	$R \times 100$	100–700
PT	AOA	$R \times 100$	100–800
PT	WWF	$R/15$	p/million

easier to compare, we added 3 units to the value, resulting in a positive range of 0.18 to 16.81 that has better legibility.

There was a need to connect each word from one language to another. This connection was performed by translating each English (EN) word to Portuguese (PT) and Spanish (SP), and each Spanish word to Portuguese and English. For this task a Perl Module, `Lingua::Translate` , was used with a backend for Google's translation system .

There are two approaches for word connection, each resulting in different levels of confidence. The first method consists in linking words from each language through its equivalent in Portuguese (translating each word to Portuguese and using that word as the pivot element).

Method 1 (M1) { Translate-To-PT(EN Word) = PT Word; Translate-To-PT(SP Word) = PT Word

The second method consists primarily in matching Spanish and English words with a single equivalent word in Portuguese, and also ensuring that English-to-Spanish and Spanish-to-English translations match.

Method 2 (M2) { Translate-To-PT(EN Word) = PT Word; Translate-To-PT(SP Word) = PT Word; Translate-To-EN(SP Word) = EN Word; Translate-To-SP(EN Word) = SP Word

This last method will prevent duplicate words and result in more accurate translations, though it will render a smaller intersection set.

The last step consists in computing statistics, including amplitude and mean for each language, and correlations and mean of absolute differences between values for each pair of languages.

4. Results and Discussion

The reliability of each statistic in English-Portuguese, Spanish-Portuguese and English-Spanish triangulation was analyzed in two different stages:

- **Stage 1:** the distance mean for each statistic and language pair was computed, i.e., the mean of the absolute values of the difference between values of a given statistic in a language pair.
- **Stage 2:** correlation (ranging from -1 to 1) was computed for each pair of languages. A value close to 0 shows that there is no relationship within the variables, whereas a value close to $+1$ or -1 indicates that the variables are related.

4.1. English–Portuguese and Spanish–Portuguese with M1

Considering English and Spanish words connected only unidirectionally to Portuguese words, as shown in Table 3, correlations for IMG have high values (0.78 and 0.70), with an average of absolute differences of 15.16% from English and 15.86% from Spanish, thus suggesting a more viable triangulation with an average error of 15-16%. In plain terms, an error of 16% on a 7-point scale corresponds to an error of 1 point, which translates to the minimum possible error, one likely to occur when asked of participants to estimate a word's statistic.

Although AoA also has high correlations (0.62 and 0.80), its smaller sample of only 78 words from Spanish may well be viewed as less reliable. On the other hand, its sample of 307 words from English appears less unreliable suggesting a triangulation with an average error of 11.51%.

Table 3: Differences between EN–PT & ES–PT values: Language Pair (LP), Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 1.

Statistic	LP	N	M	A	C
FAM	EN-PT	457	14.52	0.00–53.83	0.29
FAM	ES-PT	536	15.71	0.00–70.67	0.18
IMG	EN-PT	213	15.16	0.17–58.00	0.78
IMG	ES-PT	248	15.86	0.00–72.83	0.70
AOA	EN-PT	307	11.51	0.17–66.17	0.62
AOA	ES-PT	78	7.86	0.33–28.33	0.80
LOG-WWF	EN-PT	863	8.13	0.00–40.90	0.54
LOG-WWF	ES-PT	1350	13.00	0.00–47.86	0.42

4.2. English–Portuguese and Spanish–Portuguese with M2

When connecting English and Spanish words bidirectionally to Portuguese (see table 4), a slight reduction of connected words occurs. Correlations for IMG have small but significant increases (0.78 to 0.86 and 0.70 to 0.83). For AoA there is a minor increase (0.80 to 0.82) in Spanish to Portuguese, which is meaningful (0.62 to 0.79) in English to Portuguese, ensuing a more confident triangulation with an even better average error of 8.9%. Lastly there is a major increase in LOG-WWF (0.54 to 0.81 and 0.42 to 0.85), adding one more statistic to the list of possible reliable triangulations with smaller average errors of 4.68% and 4.91%.

4.3. English–Spanish with M1

When comparing English and Spanish statistics, connected by their Portuguese word equivalent, a larger pool of words becomes available, along with their corresponding statistics. This may well result in a more confident analysis, reflected in Table 5. After careful examination, IMG once again tends to be a more reliable triangulation with a high correlation of 0.62 and an average error of 12.88%.

4.4. English–Spanish with M2

A decrease in connected words occurs when connecting English to Spanish by the Portuguese translation, and also by ensuring that the English to Spanish translation matches the original Spanish word, and that the Spanish to English translation matches the original English word. Despite this relatively large

Table 4: Differences between EN–PT & ES–PT values: Language Pair (LP), Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 2.

Statistic	LP	N	M	A	C
FAM	EN-PT	340	13.21	0.00–38.50	0.40
FAM	ES-PT	292	14.38	0.00–57.50	0.30
IMG	EN-PT	152	15.60	0.17–58.00	0.86
IMG	ES-PT	138	14.99	0.00–72.83	0.83
AOA	EN-PT	227	8.90	0.17–49.50	0.79
AOA	ES-PT	65	7.88	0.33–28.33	0.82
LOG-WWF	EN-PT	505	4.68	0.00–27.82	0.81
LOG-WWF	ES-PT	445	4.91	0.00–23.54	0.85

Table 5: Differences between EN & ES values: Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 1.

Statistic	N	M	A	C
FAM	4 030	13.17	0.00–72.67	0.35
IMG	3 950	12.88	0.00–75.00	0.62
AOA	86	7.67	0.00–52.33	0.44
LOG-WWF	32900	10.34	0.00–62.54	0.40

decrease, their absolute values are still very high. As shown in Table 6, IMG increases considerably in its correlation (0.62 to 0.73), supporting the hypothesis that this is a reliable statistic for triangulation with a slightly smaller average error of 11.01%. Another increase occurs in LOG-WWF (0.40 to 0.79) once again confirming this statistic's possible triangulation with an even smaller average error of 5.33%.

Table 6: Differences between EN & ES values: Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 2.

Statistic	N	M	A	C
FAM	1 720	11.32	0.00–59.17	0.60
IMG	1 684	11.01	0.00–75.00	0.73
AOA	58	5.53	0.17–35.17	0.66
LOG-WWF	7 651	5.33	0.00–36.98	0.79

5. Conclusions

This paper reports a research on *Automating psycholinguistic statistics computation based on other languages' statistics*, including an analysis of the English program called N-Watch, the Spanish program called B_PAL and, lastly, a smaller Portuguese database.The analysis of these tools and of the available Portuguese databases (eg. CORLEX, PORLEX), acknowledges the need to design and implement a similar tool that can incorporate existing databases and promote research in Psycholinguistics for the Portuguese language.

For this paper's research Psycholinguistic statistics were imported from those three databases, including Familiarity, Imageability, Age of Acquisition and Written Word Frequency.

Although higher pools of words are available by connecting words only with their Portuguese translation, these words include erroneous or duplicate translations, which results in less reliable data. To enhance this, a stronger approach was used to enable a more accurate connection between each language's words. Although resulting in a smaller pool of words and statistics, these results appear more reliable allowing for a more confident and accurate inference of possible triangulations.

Results for correlation and mean of absolute differences between each language's values for each statistic seems to indicate that Imageability and WWF (through the analysis of LOG-WWF) may be automatically computed with some reliability from a triangulation from English and Spanish languages. Age of Acquisition yield interesting results, although further analysis with larger samples will be needed to conclude about their reliability.

6. Acknowledgments

This paper stems from project "Procura-Palavras (P-Pal): a software program for deriving objective and subjective psycholinguistic indices for European Portuguese words" (PTDC/PSI-PCO/104679/2008), funded by FCT (Foundation for Science and Technology), by NSRF (National Strategic Reference Framework), and by COMPETE (Operational Agenda for Competitiveness Factors), a program created by the EU as part of the European Regional Development Fund (ERDF).

7. References

[1] C. J. Davis, "N-Watch: a program for deriving neighborhood size and other psycholinguistic statistics," *Behavior Research Methods*, vol. 37, no. 1, pp. 65–70, 2005.

[2] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database (release 2)[cd-rom]," *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*, 1995.

[3] C. J. Davis and M. Perea, "BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in spanish," *Behavior Research Methods*, vol. 37, no. 4, pp. 665–671, 2005.

[4] N. Sebastián-Gallés, M. A. Mart, M. Carreiras, and F. Cuetos, "LEXESP: léxico informatizado del espanol," *Barcelona: Edicions de la Universitat de Barcelona*, 2000.

[5] I. Gomes and S. L. Castro, "Porlex, a lexical database in european portuguese," *Psychologica*, vol. 32, p. 91–108, 2003.

[6] J. Marques, F. Fonseca, A. Morais, and I. Pinto, "Estimated age of acquisition norms for 834 portuguese nouns and their relation with other psycholinguistic variables," *Behavior Research Methods*, vol. 39, no. 3, pp. 439–444, Aug. 2007.

[7] M. Nascimento, J. Casteleiro, M. Marques, F. Barreto, and R. Amaro, "Léxico multifuncional computorizado do português contemporâneo [Multifunctional computational lexicon of contemporary Portuguese](data file)," *Available from Centro de Linguística da Universidade de Lisboa Web site: http://www.clul.ul.pt*, 2000.

[8] M. F. B. do Nascimento, J. M. Casteleiro, M. L. G. Marques, F. Barreto, R. Amaro, and R. Veloso, "Corlex: Léxico de frequências do português [Base lexical]," *Consultado em http://www.clul.ul.pt (30 de Outubro de 2007)*, 2007.