Quantitative Comparison of Homonymy in Spanish EuroWordNet and Traditional Dictionaries*

Igor A. Bolshakov, Sofia N. Galicia-Haro, Alexander Gelbukh
Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
{igor,sofia,gelbukh}@cic.ipn.mx

Abstract. A quantitative comparative study of homonymy in four well-known electronic Spanish dictionaries—EuroWordNet and three traditional dictionaries—is presented. It is shown that though structuring of word senses is quite different in all dictionaries under comparison, EuroWordNet differs from the traditional dictionaries much more than these differ from each other. It is also shown that the ordering of the word senses in Spanish EuroWordNet less agrees with the use of the senses in texts than the ordering in traditional dictionaries.

1 Introduction

Different dictionaries usually give different sense sets for the same words. In this work we present quantitative evaluation and comparison of word sense structuring in the following four well-known Spanish electronic dictionaries: of Anaya group [1], by María Moliner [2], of Spanish Royal Academy [3], and EuroWordNet [4]. Our motivation was to proof or disproof the following assumptions:

- The dictionaries tend to have similar sense sets, since[1] (1) all good lexicographers share the same word sense structures in their minds, and (2) if a lexicographer does not elaborate the sense structure for a given word, he or she borrows some parts of it from other dictionaries.
- EuroWordNet dictionaries (in particular, Spanish) have made some disruption in the lexicographic tradition since they were compiled on a different ideological basis—by computer-oriented linguists and without deep lexicographic considerations.

Our motivation was also to check whether simple statistical methods could be useful for selecting a 'better' dictionary for future applications.

2 Comparison of the Dictionaries

Experimental setting. Ideally, the comparison methods discussed below operate on the representation of the dictionaries as very large sets of ordered lists (word senses

[*] Work done under partial support of CONACyT, SNI, and CGEPI-IPN, Mexico.

[1] I. Mel'čuk, private communication.

for each word) and the mappings between these lists (the correspondences between the word senses in different dictionaries); what is more, one of our experiments would, ideally, rely on the textual frequencies of specific word senses. However, given the large amount of senses in all four dictionaries, constructing such mappings and counting the frequency of each sense would be too expensive.

To simplify our calculations, we worked with small randomly chosen samples of the dictionaries. Though we realize that our results are then quite approximate, we believe they do show the general tendencies.

First, we constructed a small corpus marked with senses. We started from the well-known LEXESP corpus,[2] which contains a balanced representation of modern Spanish and has the size of 5 million words. Of those, we have randomly (by the position in the file) chosen 158 words and, basing on the context, assigned them the senses from all four dictionaries.

Then, to further simplify our calculations, we eliminated some words from this corpus: (1) In two cases, we eliminated words with the same sense, so that all words in our toy corpus had different senses. Since there were only few repeated senses, this should not affect the results but simplifies our calculations. (2) We also eliminated the words that could not be assigned a sense in at least one dictionary; there were 27% of such words, the majority of them being adjectives absent in EuroWordNet.

After these operations, we obtained a corpus of K = 114 supposedly most frequently used word senses, marked each one with a word sense number according to each of the four dictionaries. A fragment of the complete list of words is presented in Appendix 1.

This, in turn, is equivalent to the selection of a small sample of each of the four dictionaries, reflecting mainly the most frequently used senses. All our calculations described below are based on these samples instead of complete dictionaries.

Comparison of the number of word senses. For each word (letter string) w of our corpus, we found the number x_{wd} of its senses in each dictionary d. The values x_{wd} distributed as follows:

	Anaya	Moliner	Academy	WordNet
Average $\bar{x}_d$	5	4.5	7.3	3.6
Median	4	3	5	2.5

where

$$\bar{x}_d = \frac{1}{K}\sum_{w=1}^{K} x_{wd}\ .$$

It can be seen that EuroWordNet has considerably less senses per headword.

The similarity between the numbers of the senses in the entries of the dictionaries d_1 y d_2 calculated by Pearson's formula [5]

$$P_{xy} = \frac{\frac{1}{K}\sum_{i=1}^{K} x_i y_i - \bar{x}\,\bar{y}}{\sqrt{\left(\frac{1}{K}\sum_{i=1}^{K} x_i^2 - \bar{x}^2\right)\left(\frac{1}{K}\sum_{i=1}^{K} y_i^2 - \bar{y}^2\right)}}$$

[2] Kindly provided to us by H. Rodríguez of Universitat Politècnica de Catalunya.

where $\bar{x} = \bar{x}_{d_1}$, $\bar{y} = \bar{x}_{d_2}$, $x_i = x_{id_1}$, $y_i = x_{id_2}$, is as follows:

	Anaya	Moliner	Academy	WordNet
Anaya	1.000	0.812	0.947	0.565
Moliner		1.000	0.826	0.616
Academy			1.000	0.556
WordNet				1.000

It can be observed that the correlation between EuroWordNet and the other dictionaries is smaller than among these three.

Comparison of the ordering of senses. Using our toy corpus, we compared the positions of the senses within their groups for a given word (letter string) in the four dictionaries.

Let i = 1, ..., K be the number of word in our corpus, x the number of dictionary, and k_{ix} = 1, ..., n_{ix} the corresponding sense number out of n_{ix} senses in total for the corresponding letter string in the corresponding dictionary. Then the relative position

$$r_{ix} = \begin{cases} \dfrac{k_{ix} - 1}{n_{ix} - 1}, & \text{if } n_{ix} \neq 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

reflects how far the given sense is from the top of the list of senses for the given word; note that if $k_{ix} = 1$ then the relative position is 0 independently of the total number of senses n_{ix}. Note also that in the case $n_{ix} = 1$ it always holds $k_{ix} = 1$, thus the second option in (1). So we calculate the mean ordering distance between the dictionaries x and y as:

$$D_{xy} = \frac{1}{K} \sum_{i=1}^{K} \left| r_{ix} - r_{iy} \right|$$

The obtained values of D_{xy} are as follows:

	Anaya	Moliner	Academy	WordNet
Anaya	0.000	0.207	0.167	0.386
Moliner		0.000	0.254	0.388
Academy			0.000	0.411
WordNet				0.000

Once more, EuroWordNet dictionary differs from the other three considerably more that these three from each other.

Suitability of the ordering of senses. We expect that the lexicographer should list first the (intuitively) most frequent senses. Thus, using our toy corpus of (supposedly) most frequent senses, we considered the distribution of the relative positions of these senses calculated by the formula (1), which proved to be the following:

	Anaya	Moliner	Academy	WordNet
Average ($\bar{x}_d$)	0.271	0.164	0.314	0.419
Median	0.000	0.000	0.184	0.310

As one can see, the ordering of senses agrees very well with the frequencies of usage for Anaya and Moliner dictionaries. For Academy dictionary, the agreement is slightly less probably because it contains many obsolete senses. Finally, the ordering of senses in the EuroWordNet dictionary seems to be close to random.

3 Conclusions

All four dictionaries under comparison are different both in the mean number of senses per word (letter string) and in their ordering of senses for a given word. Hence, our first assumption can be rather rejected. We can admit, however, that a deeper lexicographic research can show whether these differences are mainly due to very infrequent senses, such as dialectic. Spanish is spoken by almost 400 millions of people in many countries that have great dialectic differences.

The three traditional dictionaries have greater differences with EuroWordNet than between each other. Specifically, Spanish EuroWordNet has significantly less number of senses, lacking quite frequently used senses (especially adjectives). While the ordering of senses in the traditional dictionaries agrees quite well with the relative frequencies of their usage, the ordering of EuroWordNet seems to be almost random. Thus, our second assumption has been confirmed.

References

1. Anaya Group. Diccionario Anaya de la lengua. Internet. Marzo 1997.
2. María Moliner. Diccionario de uso del español. GREDOS Primera edición en CD-ROM 1996.
3. Real Academia Española. Diccionario de la Lengua Española. Edición vigésima primera, en CD-ROM de ESPASA CALPE. 1995.
4. EuroWordNet Consortium, Spanish version. 1997-1998.
5. McEnery, T. & A. Wilson. Corpus Linguistics. Edinburgh University Press. 1996.

Appendix 1. Examples of homonyms in the text we investigated

In the table below, the number k_{ix} of the sense in our corpus and the total number n_{ix} of senses for the given word (as letter string) are given. Here (N) stands for noun, (A) for adjective; unmarked words are verbs.

Word (Spanish)	English	Anaya	Moliner	Academia	EWnet
aceleración N	acceleration	1/2	1/3	1/2	2/2
alcanzar	to reach	4/8	4/8	7/18	2/4
año N	year	2/3	1/3	3/7	2/3
apresurar	to hasten	1/2	1/2	1/2	2/4
asunto N	affair	1/4	1/2	6/6	6/6
atención N	attention	1/2	1/5	1/4	5/7

comida N	dinner	3/4	3/3	2/4	7/8
creer	believe	1/6	1/3	1/5	2/6
dar$_1$	overlook	24/29	10/12	38/47	9/9
dar$_2$	to cause	7/29	4/12	21/47	6/8
decir	to say	1/8	1/10	1/10	3/8
diario N	newspaper	2/3	2/5	4/5	2/6
dormir	sleep	1/6	1/8	1/12	1/2
encontrar	to meet	3/8	2/4	5/8	9/9
enseñar	point out	2/6	2/3	3/6	4/8
girar	to turn	1/5	1/7	1/7	7/15
hombre$_1$ N	male	2/5	1/4	2/10	4/6
hombre$_2$ N	adult	3/5	1/4	3/10	4/6
llamar	to name	7/12	4/10	4/13	8/12
mover	to move	1/9	1/9	1/10	2/2
nido N	nest	1/4	1/7	1/8	2/2
padre N	parents	5/8	7/9	10/12	1/2
pasar	go through	3/35	9/41	24/59	19/21
posible A	possible	1/2	1/2	1/2	1/2
proyecto N	plan	2/3	1/2	4/4	1/3
régimen N	government	2/6	2/3	2/7	3/3
rendimiento N	income	1/3	1/2	4/5	4/4
situación N	situation	2/3	2/2	4/6	4/7
tener	possess	2/13	2/13	2/24	4/4
terreno N	field	4/6	4/5	3/6	4/4
varón N	male	2/3	2/3	2/4	1/2