The Information in General Purpose Language Dictionaries

María Paz BATTANER, Victòria ALSINA and Janet DeCESARIS,
Barcelona, Spain

Abstract
This paper summarizes a research project funded by the Spanish Ministry of Education (PB96-0305) on the information present in general purpose language dictionaries. We found that both monolingual and bilingual dictionaries were weak on pragmatic information concerning use. Entries in bilingual dictionaries were frequently not well organized in terms of order of equivalents or sense distinctions from a bilingual perspective, thus resulting in needless repetition of the same equivalent. Electronic editions of dictionaries were used to study the lexical semantics of specific affixes (the verbal prefix *re-* and the adjective suffix *-oso*), and their use was critical in establishing the semantics of the affixes. Finally, the paper discusses the representation of animal names in Spanish dictionaries and the role of encyclopedic knowledge in the definitions of these nouns.

1   Introduction

This paper reports on the work that has been done in conjunction with a grant from the Spanish Ministry of Education and Culture[1] to the Institute of Applied Linguistics at Pompeu Fabra University (UPF) in Barcelona[2]. The project studied the information contained in general language dictionaries in order to propose ways of improving the presentation of information contained in new dictionary projects.

The project concentrated on several types of headwords: adjectives, adverbs, words requiring a description involving encyclopedic information, as well as morphological units such as a specific affix. Monolingual Spanish and English dictionaries, as well as bilingual dictionaries for English and Spanish and English and Catalan were the objects of study. As will be seen from the discussion, both printed and CD-ROM editions of dictionaries were consulted.

Although it is fair to say that Spain is a country with a strong lexicographic tradition, it is only quite recently that the information present in dictionary entries has been the object of systematic analysis. In this summary we present an overview of several of the issues studied in the course of the past three years. In order to give the reader an overall view of the project, we will not dwell on specific examples, nor are we able to discuss all the points we have touched on. Our goals here are to identify the areas in which the information provided by the dictionary was satisfactory, and those in which the information needs to be improved or substantially restructured in order to be of maximum use to the target audience.

2   Specific points studied in the project

2.1   Adjectives

Battaner first noticed the frequent use of *debido*/*debida* 'due' while studying the definitions of verbs beginning with the prefix *des-* 'de-' in the *Diccionario de la lengua española de la Real*

Academia Española (DRAE 1992). Examples of *debido*/*debida*, along with those of the related adverb *debidamente*, were taken from the electronic editions of three monolingual dictionaries; the forms were present in 377 definitions[3]. *Debido* is an adjective that establishes a relationship in terms of a standard that is often social in nature, and is used by Spanish lexicographers in contexts in which a social norm is recognized to exist. Battaner questions whether dictionaries should include this sort of value judgment in definitions. The paper notes that *debido* is the past participle form expressing a value that is generally accepted by society, and shows that it cannot be replaced in certain definitions by more neutral modifiers; in other words, certain words can only be defined with reference to social norms. In this sort of context, the agent of the action *deber* is unknown and/or so distant as to be unidentifiable, which is exactly the case for the authors of social norms. Interestingly, comparison of the three dictionaries shows that the dictionaries seldom coincide in terms of which definitions contain this marker (all three coincide in only 7 cases, two of the three in 13 and 9 cases), despite the fact that all these dictionaries were written and published in Spain.

Alsina and DeCesaris have analyzed the treatment of several types of adjectives in monolingual and bilingual dictionaries in a series of papers[4]. Their work on equivalents in bilingual dictionaries has focused on entries for non-derived, polysemous adjectives for which bilingual dictionaries provide several equivalents. The choice among equivalents is an advantage for users who are practically bilingual, such as professional translators, for whom the dictionary often provides confirmation of already known information. But for advanced language learners, one of the main target groups of large bilingual dictionaries, a choice among equivalents with little or no contextual information and no examples is not a choice; it is simply a lot of (potentially useful) information. In their analyses of one Catalan-English and three Spanish-English bilingual dictionaries[5], they found a wide variety of approaches to the problem of providing equivalents for polysemous adjectives.

While one dictionary contains entries with no apparent internal organization but many translation equivalents for fixed expressions, another dictionary may be much better structured but in the end provide less information to users because fewer examples are included. Their research confirms that in general, the more recent a bilingual dictionary, the better organized the entries are, especially in terms of providing important pragmatic and grammatical information about adjectives. This was particularly true for indications concerning adjective placement in Spanish; the most recent dictionary contained clearly presented information concerning differences in meaning, whereas the same information was generally present but not easily recoverable in older dictionaries.

Examination of entries for adjectives shows that bilingual dictionaries of English and Spanish or Catalan and English are not conceived of as contrastive works, but rather are based on one language, with equivalents in the other being listed to varying degrees of completeness. This approach often leads to excessive repetition of a single equivalent (e.g. the Sp. equivalent for Eng. *high*, *alto*, may be given in several places in an entry.), or overt representation of morpholexical information that may be unnecessary given the expected target audience of the dictionary (e.g. repetition of gender endings for all Spanish adjectives in both directions). Even though the worldwide market for general bilingual dictionaries aimed at native speakers of Spanish and English is huge, analysis of the entries shows that publishers continue to try to address issues of interest to native speakers of both languages in both directions of the dictionary, and the results

are rarely satisfactory for either group of speakers.

Two other problematic issues were discussed in relation to these same bilingual dictionaries, namely: (1) determining the proper role for translated phrases and expressions; and, (2) identifying the proper way to order the equivalents presented. The fact that bilingual dictionaries often do not clearly distinguish between a separate sense and the specific translation equivalent for a fixed expression may result in more sense distinctions being present in a bilingual dictionary than in a monolingual dictionary aimed at educated native speakers; e.g. for the adjective *high*, *The American Heritage Dictionary* (1992) lists 13 main senses, with a total of 22 sub-senses, whereas the bilingual *Larousse Gran Diccionario* (1983) lists 27 senses. The authors note that this should not be the case, since general monolingual dictionaries cover uncommon and archaic senses that rarely play a role in bilingual lexicography.

The bilingual dictionaries analyzed differ greatly in terms of the order of equivalents presented (although the equivalents and expressions themselves often vary little), both in comparison with each other and when compared with monolingual dictionaries. Better organization of entries in bilingual dictionaries, particularly in terms of placement of translation equivalents of phrases, is urgently needed for these two language combinations.

2.2 Words with a specific affix

Battaner used the CD-ROM version of *DRAE 1992* to analyze the semantic relationships established by the prefix *re-*, to see if the definitions of the relevant verbs could help define exactly what information *re-* added to a verb[6]. This prefix is both interesting yet problematic because it is present in both popular and learned words as well as in borrowings, thus potentially making it more difficult for speakers to associate it with a specific meaning. Analysis of the definitions shows that verbs with *re-* form a part of a complex lexical semantic network. The scope of this prefix covers (from least frequent to most common meaning) opposition to, regression from, progression towards, repetition of, and intensity (quantitatively and qualitatively). Use of the CD-ROM version of the dictionary proved essential in determining the possible range of meanings, as verbs could be grouped together by keywords found in the definitions.

Alsina and DeCesaris also looked at the relationship between morphology and lexicographic representation in monolingual dictionaries. The differences in the morphological characterization of the suffixes *-ful* and *-like* in English were well reflected in the entries in English monolingual dictionaries[7]. Both *-ful* and *-like* historically derive from words, but accepted morphological analysis of English[8] places words with *-ful* in the derivational component whereas words with *-like* are compounds. The electronic edition of *Merriam-Webster's Collegiate Dictionary* was consulted and analysis showed that words with *-ful* are more highly lexicalized and thus afforded definitions, whereas words with *-like* are more often listed as run-on entries.

Given that the above research supported the view that differences in morphological structure are evident in dictionary representation, the authors then posed the question of how information present in dictionaries could be used to resolve problems in morphological analysis. Most research in morphology over the past thirty years has assumed that a single suffix does not attach to more than one morphosyntactic category[9]. The Spanish adjective suffix *-oso* attaches to nouns, verbs, and adjectives, and thus under the widely accepted view in morphology there

must be at least two suffixes, one that attaches to nouns and adjectives and one that attaches to verbs. All the words containing the suffix beginning with the letters B and C listed in the CD-ROM version of the *DRAE* were analyzed to see if the information in the dictionary supported this morphological analysis[10]. Spanish lexicographic tradition, of which the *DRAE* is the prime example, does not employ run-on entries, so that all these derived adjectives had definitions. The suffix *-oso* is defined in the *DRAE* as indicating presence of a characteristic or substance, abundance thereof, or tending to exhibit characteristics similar to those associated with the base without actually obtaining them; e.g., *verde - verdoso* 'green - greenish'. The dictionary thus states that the suffix is polysemous, but makes no comment on the morphosyntactic nature of the base. Using the definitions in the *DRAE*, the authors argue that *-oso* is a single, polysemous suffix and that its core meaning is presence of the base, and show that it attaches to many more nouns than verbs. The idea of 'abundance of', which is commonplace in the definitions of the *DRAE*, reflects the fact that pragmatically one only uses a property to characterize a noun when that property is salient, which usually means it is present in abundance (e.g. fried food is *aceitoso* 'oily' only when there is too much oil so as to be worthy of comment, not simply when some oil is present).

2.3 Adverbs

Martínez noted that although adverbs and discourse markers have been the subject of much interest in Spanish language studies in the past 25 years, Spanish dictionaries have, by and large, not incorporated the pragmatic information associated with these expressions. This is perhaps surprising given that Spanish lexicographic tradition has long incorporated information concerning usage and that in Spain there is often a close working relationship between grammarians and lexicographers in university as well as professional settings. Adverbs are often defined with synonyms or periphrases, but the entries rarely contain any examples or information about how the words are used in context. Comparison of entries from three general monolingual Spanish dictionaries[11] shows that these expressions do not receive homogeneous treatment, particularly in comparison to the type of entries used for other classes of words. One reason which might explain this lack of uniformity in approach is that the dictionaries analyzed were written almost exclusively as tools for comprehension. Dictionaries need to include more semantic and pragmatic information for these expressions if they are to be of use in production tasks.

2.4 Definitions involving encyclopedic descriptions

Borràs focused her attention on the information contained in the entries for names of animals. She consulted the CD-ROM editions of three comprehensive, general language dictionaries[12] in order to determine what information Spanish lexicography has traditionally included about this part of the lexicon. Because these headwords usually have encyclopedic definitions, it is important to determine what kind of extralinguistic information appears in the definition, and whether this information is sufficient or not. Definitions of animals can change over time as a result of progress in the field of biology, but at times these changes conflict with popular beliefs. This has significant repercussions for the descriptor used to classify the noun in the definition. Interestingly, the three dictionaries analyzed do not classify animals in the same way, and there are 'errors' in classification that can be accounted for by taking anthropological and

cultural attitudes into account. The database developed is also interesting from the standpoint of analyzing figurative vs. literal meaning in dictionary definitions, as these nouns are frequently the basis for lexicalized or semi-lexicalized metaphorical interpretation.

3 Discussion and implication for future study

It is hardly surprising that the information contained in large, general purpose language dictionaries is extremely wide in scope and nature, but that does not mean we know everything about what these dictionaries contain. It is not only a matter of identifying what is present and what is absent, with a view to advancing proposals to fill gaps, but also a matter of uncovering the information that lies behind the definitions. Analysis of dictionary entries using computerized technology allows us to form a more accurate picture of the information because computer searches are systematic. We are thus very encouraged by the results of using electronic editions of dictionaries for academic research on lexicography; we note, however, that not all Spanish commercial publishers are convinced of the need for general purpose dictionaries on CD-ROM.

The ability to search for specific phrasing in definitions allowed us to uncover an ideological factor in the use of *debido* 'due' in reference to established social norms. That factor would have been likely to remain unnoticed had it not been for the ability to search for all definitions containing the adjective in question. Computer searches also allowed us to scrutinize the meaning of words with a single formant. The semantics of morphologically derived words is seldom directly compositional, and work done in this project shows how lexicographical definitions can be used to aid in determining the meaning of a particular affix.

The Spanish dictionaries studied were found to be weak on discourse-related information. There has been a tendency in the two last editions of the most influential dictionary in Spanish lexicography [DRAE 1984, DRAE 1992] to delete derived adverbs from the list of headwords defined in the dictionary. This practice would seem to deny the role that adverbs like *probablemente* 'probably' or *seguramente* 'surely' play in discourse. In this case, then, the dictionaries studied did not supply enough information for users. In contrast, the information found in the bilingual dictionaries studied was generally adequate but often buried in poorly organized entries. Spanish/English bilingual lexicography does not always draw a clear line between sense distinction in equivalents and translation equivalents for set phrases, which results in long, repetitive entries that contain many translated expressions, all of which have the same equivalent. It is not clear to us that these long articles are necessary; the same number of useful examples could occupy less space if the entry were better organized. The grammatical information needed in a bilingual dictionary particularly depends to a great extent on the native language of the user, and the commonplace practice of publishers of aiming to please everyone all the time results in bilingual dictionaries that are always lacking some information for some group; dictionaries designed for more specific target audiences would seem preferable, and in the case of Spanish and English, are financially viable projects. Although [Urdang 2000] states that such an approach is widely put into practice, at least for some language combinations at some publishers, we found little evidence of user perspective in the bilingual dictionaries studied.

The encyclopedic information used to describe a single class of nouns (that of animals) varies greatly from dictionary to dictionary. Spanish lexicography traditionally distinguishes between

'language' dictionaries and 'encyclopedic' dictionaries. Some encyclopedic information is necessary in order to interpret lexicalized, or semi-lexicalized, figurative uses of animal names, which are often listed as separate senses, but exactly how much real world knowledge should be used to define these nouns in dictionaries that aim to minimize the role of this sort of information is an unresolved issue.

One of our main objectives at the start of this project was to determine if the dictionaries under study needed to provide more information; i.e. were the problems that we experience as language and translation teachers due to a lack of information available to users. We have seen that some but not all issues are quantitative. A question that arose in much of our work was how to deal with lexicalized metaphorical interpretation of words and expressions requiring pragmatic and encyclopedic knowledge. We hope to continue working on that and other issues in the future, in order to help Spanish and Catalan lexicographers to become fully aware of the scope of the information included in their dictionaries.

Notes

[1]Grant PB96-0305 from Ministerio de Educación y Cultura of Spain to the Institut Universitari de Lingüística Aplícada (María Paz Battaner, principial researcher), Universitat Pompeu Fabra is hereby gratefully acknowledged.

[2]The lexicography group works under the direction of María Paz Battaner, and includes eight researchers, four of whom hold doctorate degrees (Battaner, Alsina, DeCesaris and Christina Gelpi) and four of whom are currently enrolled in the Ph.D. program in Applied Linguistics at the UPF (Laura Borràs, Núria Castillo, Roser Martinez and Sergi Torner).

[3]The dictionaries studied were the *Diccionario de la lengua española de la Real Academia Española* (21st edition), 1992; CD-ROM 1995), the *Diccionario de Uso del Español* by María Moliner (Gredos, Madrid, CD-ROM 1996) and the *Diccionario General de la Lengua Española* (Bibliograf, Barcelona, CD-ROM, 1997). The adjective form was found in 328 definitions, and the adverb in 49. The work discussed here was presented at the Romance Philology Seminar of the University of Augsburg in 1997.

[4]The papers are 'Definiciones de adjetivos en diccionarios monolingües y bilingües.' *Los distintos dominios de la Lingüística Aplicada desde la perspectiva de la Pragmática* (Ignacio Vázquez Orta and Ignacio Guillén Galve, eds.) Zaragoza, Spain: ANUBAR ediciones, 1998, pp. 347-353; 'La lexicografía bilingüe y la colocación del adjetivo: problemas de equivalencia,' paper read at the Annual Meeting of the Spanish Applied Linguistics Association, Logroño, 1998, to appear in the proceedings; 'Bilingual lexicography, overlapping polysemy and corpus use,' paper read at the 1999 Louvain la Neuve Conference on Translation, Bilingual Lexicography and Corpus Use, to appear in *Languages in Contrast* ed. by B. Altenberg and S. Granger, John Benjamins, Amsterdam; '*Todo y cada* y la equivalencia bilingüe castellana-inglesa,' paper read at the III Congress of General Linguistics, Salamanca, 1998, to appear in the proceedings; 'Dues llengües, dues tradicions: els diccionaris catalans i americans,' paper read at the VII Colloquium of the North American Catalan Society, Barcelona, 1999, to appear in the proceedings.

[5]The dictionaries studied were the *Oxford Spanish Dictionary* (1994), *Simon and Schuster's International Dictionary Spanish-English/English-Spanish* (1971), and *Diccionario Gran Larousse Español-Ingles/English-Spanish* (1983).

[6]Cf. Battaner, María Paz. 'Estructura del vocabulario a través de las definiciones lexicográficas: Las entradas *re-* del DRAE 1992' *Filología e informatica: Nuevas tecnologías en los estudios filológi-*

cos (J.M. Blecua, G. Clavería, C. Sánchez and J. Torruella eds.). Bellaterra, Barcelona: Seminario del Filología e Informática, Universitàt Autonoma de Barcelona, 1999, pp. 227-257.

[7]Cf. Alsina, Victória and DeCesaris, Janet. 'Morphological structure and lexicographic definitions: the case of *-ful* and *-like*". *EURALEX '98 Proceedings* (T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin, S. Theissen eds.), University of Liège, 1998, pp. 545-554.

[8][Bauer,1983] is the standard reference on word-formation in English.

[9]This hypothesis, first advanced by [Aronoff 1976] and subsequently modified by [Scalise 1984], is known as the (Modified) Unitary Base Hypothesis.

[10]The work discussed here was presented in the paper *Los adjetivos formados con el sufijo -oso. Estudio basado en las definiciones lexicográficas*, presented at the V Simposio de Lingüistica Hispánica, Universität Leipzig, to appear in the proceedings.

[11]The dictionaries studied were the *DRAE*(1992), the *Diccionario de Uso del Español* (CD-ROM, 1996), and the *Diccionario General Illustrado de la Lengua Española* (1987).

[12]The dictionaries studied were the CD-ROM editions listed in endnote three.

References

[1] Aronoff, Mark (1976) *Word Formation in Generative Grammar*. MIT Press, Cambridge, MA.

[2] Bauer, Laurie (1983) *English Word-formation.* Cambridge University Press, Cambridge.

[3] Scalise, Sergio (1984) *Generative Morphology*. Foris Publications, Dordrecht.

[4] Urdang, Laurence (2000) Review of R. R. K. Hartmann and Gregory James, *Dictionary of Lexicography*, in *International Journal of Lexicography*, Vol. 13/1, pp. 35-42.