Improving Lexical Databases with Collocational Information: data from Portuguese

Paula Guerreiro

Centro de Linguística da Universidade de Lisboa: CLG – Computation of Lexical and Grammatical Knowledge Research Group Av. 5 de Outubro, 85 – 5°. 1000 Lisboa. Portugal Paula.Guerreiro@clul.ul.pt

Abstract

This article focuses on ongoing work done for Portuguese concerning the phenomenon of lexical co-occurrence known as collocation (cf. Cruse, 1986, *inter al.*). Instances of the syntactic variety formed by noun plus adjective have been especially observed. Collocational instances are not lexical entries, and thus should not be stored in the lexicon as multiword lexical units. Their processing can be conceived through relations linking the lexical components. Mechanisms for dealing with the collocation-hood of the expressions are required to be included in the systems, topographically, in their lexical modules. Lexical databases like wordnets, with a general architecture typically structured on semantic relations, make room for the specification of this phenomenon. This can be handled through the definition of *ad-hoc* relations expressing the different semantic effects the adjectival modification bring to nominal phrases, collocationally.

1 Introduction

This paper deals with the linguistic phenomenon called collocation and describes partial work done for Portuguese concerning related collocational instances of the syntactic variety formed by noun and adjective. Collocation is here referred to in a sense close to Cruse (1986), inter al.. Nominal phrases as jantar opíparo ('sumptuous dinner'), comunista ferrenho ('die hard communist'), sorriso amarelo ('sickly smile'), sorriso estampado ('wide smile'), pão fresco ('new bread'), peixe fresco ('fresh fish'), constitute typical examples of collocation.

The assumption of this line of research, scarcely attended in the description of the Portuguese language, has two main goals: i) to point out the linguistic relevance of the collocation concept for an adequate description of lexical co-occurrence preferences and constraints; ii) to include systematised collocational information in lexical databases, contributing in this way for the improvement of Portuguese linguistic resources.

Current perspectives of the lexicon see it as an inherently relational whole. Recent works in Psycholinguistics highlighting the effective relational architecture of the so-called mental lexicon have inspired models of lexical databases (e.g. WordNet, (Fellbaum, 1998)). Even printed dictionaries, normally sorted in alphabetical order and apparently organised on a merely enumerative basis, have demonstrated to be strongly structured upon semantic relations, as important research could evidenced (cf. Evens, 1988).

In this line of ideas, the collocation phenomenon is kept here as a kind of co-occurrence *attraction relation* linking words in expressions that cannot properly be seen as compounds. For the sake of the modelling of the specifications, such perspective leads to the description of the phenomenon by means of specific relations.

This document is structured in four sections. Next section deals with the linguistic relevance of the phenomenon. In section 3 we stress the existence of semantic productivity of collocations, and try to evaluate how to take advantage of it in an NLP perspective. Section 4 tackles the representation of collocational knowledge in lexical databases, more precisely in wordnets.

2 The linguistic relevance of collocation

Notwithstanding some indifference devoted to collocation felt in the descriptions of Portuguese, the phenomenon is apparent in some areas of applied linguistics implying discourse production, as human or automatic translation and traditional or computer assisted language learning. It gives support to the assumption that collocation must be treated as a linguistic phenomenon. Concerning NLP, the specification of collocational knowledge in lexical databases has been pointed out as being highly recommended for several applications as automatic interpretation, style and grammar checking and general generation (cf. (Calzolari, 1995:279)). We believe it would contribute in a relevant way for the linguistic realism of applications.

Despite the facts, *collocation* has not been object of an univocal conceptualisation, as it has been stressed in Heylen (1993). The concept of collocation followed in this paper has mainly a semantic/lexical relevance, as in Cruse (1986) (cf. also Hausmann, 1979 and Cop, 1991). It may not correspond to the quantified significant collocation emerging from statistical acquisition approaches (viz. Jones & Sinclair, 1971; Church &Hanks, 1990, Smadja, 1993; Kjellmer, 1994). It is worth notice that some recent works concerning high frequency Portuguese word combinations, based on quantitative approaches, do not seem to provide the suitable explanation for the particular kind of cases at stake.

2.1 Features of collocation-hood

Aspects determining collocation-hood are briefly outlined in this section. As stressed before, the perspective on collocations adopted here is anchored in a lexicographic and semantic frame. The reader can find in the related literature (cf. Benson 1985) further distinctions on the matter of collocation, namely those opposing *grammatical collocations* (generally said as recurrent co-occurrence between content words and prepositions) to *lexical collocations* (said as recurrent co-occurrence between content words). Contrarily to some authors, but in accordance to Hausmann (1979), we don't recognise here, however, this claimed 'recurrent' status of collocational instances as a definitional feature of the phenomenon, as we have pointed out in the previous section.

Collocational sequences of the variety of noun plus adjective will be observed and contrasted with free expressions presenting the same syntactic configuration. Consider the following expressions:

1.a) sapato cambado 'shoe worn at the heels'

b) sapato castanho 'brown shoe'

2.a) *ovo escalfado* 'poached egg'

b) *ovo cozido* 'boiled egg'

Both 1. and 2. provide nominal phrases syntactically well-formed. It is however easy to feel the fact that in 1.a) and in 2.a) the component lexemes co-occur in a more privileged fashion than in 1.b) and 2.b). In fact, expressions 1.a) and 2.a) constitute instances of collocation, seen as a phenomenon of privileged or even exclusive (cf. 2.a) lexical co-occurrence in languages, from which result well-formed sequences. Lexical items entering into one collocational sequence seem to be mutually selective, maintaining what can be informally said as an 'attraction', although it could be possible, in semantic grounds, to capture the polarity of the selection (in a head-modifier construction, the selector corresponds to the modifier, which presupposes one or more semantic traits to its selectee (cf. Cruse 1986:104)).

Consider now sequences 3. and 4.:

3. a) sorriso estampado 'wide smile'b) tecido estampado 'patterned cloth'

4. a) sorriso amarelo'sickly smile'b) tecido amarelo'yellow cloth'

Here, the expressions 3.a) and 4.a) exhibit an anomalous and unpredictable meaning, which is not easily analysable through compositional processing, as it is for 3.b) and 4.b). The semantic abnormality of expressions 3.a) and 4.a) is clearly imposed by the adjectives, and these combinations are said to be xenonyms, being formed by dissonant components. The compositional processing of such expressions fails, being afterwards rebuilt in a parallel modulation, which can be called idiomatic or suppletive processing.

In spite of this 'semantic pathology', the resulting meaning of the sequence 3.a) remains transparent. The fact that semantic transparency overrides semantic abnormality may seem paradoxical or, at least, somehow mysterious. It is probably the usual way of formation of figurativeness. Historical and cultural facts have perhaps been engraved in linguistic utterances, and fossilisation of meaning is a matter of diachrony. One may eventually argue, for instance, that the sequence *sorriso amarelo* displays the image of oriental people smiling, interpretation that associates to this colour adjective an extra geographic meaning, figurative. A diachronic perspective for the matrix of meanings, out of the scope of this text, could bring light to whether figurative senses have an homonymic or polysemic nature.

Xenonymic or idiomatic collocations rise a central issue underlying this paper. We refer to the need of collocational inventories where instances of collocation are validated as real facts of the language despite their eventual semantic abnormality. Contrast the following expressions:

Being both xenonymic expressions, 5.a) is a collocation while 5. b) is unacceptable, without real existence in the language.

3 Exploiting collocational productivity

Collocation is currently said not to be productive in the sense that it is often unpredictable. As seen in the previous section, instances may appear as bizarre or arbitrary lexical co-occurrences presenting some semantic pathology.

By productivity of collocation we mean here the possibility of existence of sets of collocational instances related by matricial features determining their collocational character. Productivity of collocation do exist, naturally weak, both in xenonymic as in philonymic cases.

Having argued that collocations must be described in the lexicon, the detection of a degree of collocational

productivity seems to be exploitable in terms of lexicon design and organisation. Looking at the expressions in 6.,

- 6. a) *lauto banquete* 'sumptuous banquet'
 - b) *lauto jantar* 'sumptuous dinner'
 - c) lauto almoço 'sumptuous lunch'
 - d) *lauta ceia* 'sumptuous supper'
 - e) *lauta refeição* 'sumptuous meal'
 - f) ?lauto pequeno-almoço 'sumptuous breakfast'
 - g) ?lauta merenda 'sumptuous snack'

we observe that the adjective *lauto* ('sumptuous') co-occur with nouns belonging to a restricted branch of an hierarchy. *Refeição* ('meal') in 6.e) is the local hyperonym, the other nouns are co-hyponyms. Note however that the adjective at stake, as selector, seems to presuppose fine-grained restrictions to the selectee nouns In 6.f) and 6.g) the two nominal co-hyponyms difficulty enter in collocation with this adjective; they are somehow incompatible with the adjective because they inherently own the property of being light.

On the other hand, in the collocational instances in 7.,

- 7. a) banha rançosa 'rancid lard'
 - b) *manteiga rançosa* 'rancid butter'
 - c) *óleo rançoso* 'rancid oil'
 - d) azeite rançoso 'rancid olive oil'
 - e) *margarina rançosa* 'rancid margarine'
 - f) gordura rançosa 'rancid fat'

all the nouns of the hyponymic series of *gordura* ('fat') are able to be modified by the adjective *rançoso* ('rancid').

The collocational productivity of the adjective can be partial or total, as shown in examples 6. and 7.. Or it can be null, existing at least, in Portuguese, one case of exclusive collocation of an adjective with a noun, referred to above in example 2. a).

The detection of the existence of collocational productivity does not mean it can be verified for all the set of nouns entering in a branch of a hierarchy. It can be economically exploited by linking an adjectival entry to a hyperonymic nominal node through a specific relation. Embedded hyponyms enter automatically, in this way, in the collocational relation applied. Exceptions can be found, however, as shown by examples in 6. Automatic

inheritance of collocational productivity can be useful in hierarchised lexicons, but must be controlled.

4 Representing collocational information in databases

Collocational sequences, frequently unnoticed in discourse decoding by native speakers, are mainly noticeable in discourse production situations. The phenomenon is especially perceived when inappropriate lexical choices affect performative naturalness. The representation of collocational information must thus be designed taking into account its usefulness for generation purposes.

Concerning the issue of the representation, emphasis is given to the fact that instances of collocations are not lexical entries and, for this reason, they should not be stored in the lexicon as multiword units, as blocks. On the other hand, it seems insufficient for their representation to dress lists, e.g. given a noun, of all the adjectives that enter into a collocational frame with. It would be little informative about the resulting semantics of a given collocational instance. Given the fact that during discourse production the speaker makes lexical choices for the encoding of the meaning he wants to express, it would be desirable that stored collocational information would be semantically structured in order to guide the speaker in his choices.

Heylen (1993) has evaluated frameworks for the encoding of collocational information. Lexical Functions (Mel'cuk 1988) devices and Qualia Structure roles (Pustejovsky 1991) appear as the best mechanisms to cope with this goal and complying with a generation oriented perspective.

Accommodating of explicit collocational knowledge in the Portuguese wordnet is being designed by means of the definition of *ad-hoc* relations linking lexemes, in the line of the proposals given in Heylen (1993). For the time being, collocation of the syntactic variety of noun plus adjective has been worked.

The relations defined, linking adjectives to nouns, express information concerning the philonymy or the idiomaticity of the collocation the lexemes enter into, the type of semantic effect the collocation imposes (by means of labels inspired in Mel'cukian Lexical Functions, cf. "MAGN" above in the example) and eventually, the specification of the particular semantic components of the noun undergoing the restricting effect of the adjective (by means of labels inspired in Pustejovksy's Qualia roles). Consider, for the sake of exemplification:

bica 'expresso' (pos:n)
is_in_philonymic_collocation_MAGN-const
 forte 'strong' (pos:adj)

peixe 'fish' (pos:n)
is_in_philonymic_collocation_BON
 fresco 'fresh' (pos:adj)

Queries performed over the adjective through the Interlingual Index link allow the capture of the corresponding instances in other languages, for a multilingual perspective on collocations .

5 Conclusions

Collocation is an effective linguistic phenomenon of privileged lexical co-occurrence. As other aspects of lexical co-occurrence, it deserves a treatment, which would contribute, in NLP, for the desired performance naturalness of applications. Collocational sequences are not lexical entries, but instances of meaningful relations between lexemes, which must be included in lexical databases.

References

- Benson, M. (1985). Collocations and idioms. In R. Ilson, Dictionaries, Lexicography and Language Learning (pp. 61-68). Oxford: Pergamon Press.
- Calzolari, N. (1995). Issues for Lexicon Building. Current Issues in Computational Linguistics: In Honour of Don Walker: Linguistica Computazionale, IX-X, 267-281.
- Church, K. W. and P. Hanks (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16 (1), 22-29.
- Cop, M. (1991). Collocations in the Bilingual Dictionary. In F. J. Hausmann et al. (Eds.), Dictionaries. Dictionaries. Wörterbucher. International

- Encyclopedia of Lexicography (pp. 2775-2778). Berlin: de Gruyter.
- Cruse, D. A. (1986). Lexical Semantics. Cambridge: Cambridge University Press.
- Evens, M. W. (1988). Relational Models of the Lexicon. Representing Knowledge in Semantic Networks. Cambridge: Cambridge University Press.
- Fellbaum, C. (Ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge/Massachussets/London: The MIT Press.
- Hausmann, F. J. (1989). Le dictionnaire de collocations.
 In F. J. Hausmann et al. (Eds.) Dictionaries.
 Dictionnaires. Wörterbucher. International
 Encyclopedia of Lexicography (pp. 1010-1019).
 Berlin/New York: de Gruyter.
- Heylen, D. (1993). Characterising collocations.Collocations and the lexicalisation of semantic information. Final Report of the ET-10/75 Project, 1, 13-34.
- Jones, S. and J. McH. Sinclair (1971). English Lexical Collocations: a Study in Computational Linguistics. Cahiers de Lexicologie, 24, 15-61.
- Kjellmer, G. (1994). Introduction. In G. Kjellmer (Ed.), A Dictionary of English Collocations, based on the Brown Corpus (pp. ix-xl). Oxford: Oxford/Clarendon Press.
- Mel'cuk, I. (1996). Lexical Functions: a Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (Ed.), Lexical Functions in Lexicography and Natural Language Processing (pp. 37-102).
 Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Pustejovsky, J. (1991). The Generative Lexicon. Computational Linguistics, 17 (4), 409-441.
- Smadja, F. (1993). Retrieving collocations from text: Xtract. Computational Linguistics, 19 (1), 143-177.