The Preposition DE in Brazilian Portuguese as a Verbal Link

Aline Villavicencio[1] and Maria José Finatto[2]

[1] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
Department of Computer Sciences, Bath University (UK)
e-mail: avillavicencio@inf.ufrgs.br
[2] Institute of Language and Linguistics, Federal University of Rio Grande do Sul (Brazil)
e-mail: mfinatto@terra.com.br

Abstract. The preposition *de* is the most frequent word in written text in (Brazilian) Portuguese. However, investigation of its behavior and distributional properties is lacking. Such an understanding about as frequent words as *de* is crucial for improving the results of many NLP tasks and applications such as Machine Translation. In this paper we attempt to move one step forward in the profiling of this word, by investigating its description in linguistic resources and its occurrence in corpora, focusing on its use with verbs.

1 Introduction

In written text the preposition *de* (of) is the most frequent word used in Brazilian Portuguese. However, specific studies are lacking about the properties of this word, even though knowledge about the usage and grammatical distribution of words as frequent as this one is important for Natural Language Processing and related disciplines.

The importance of understanding the properties and behavior of words as frequent as *de* has long been recognized. For instance, Bame [1] investigates resultative and accusative patterns of the preposition up, one of the most frequent in English, while Abeillé et. al. [2] investigate the behavior of two very common French words: the prepositions *à* and *de*. Studies like these can inform NLP tasks and applications, such as Machine Translation, and lead to a more accurate characterization of languages and better results. A better understanding of *de*, with its linking function, can contribute, for example, to the identification of terms and other multiword expressions (MWEs) [3]. Dias and Lopes [4], in a study of a 300,000 words Portuguese corpus, find among the MWEs a predominance of conjunctive or prepositional expressions, such as *por parte de* (from), and many contain *de*. They also remark that their frequency and diversity in corpus is much larger than suggested in grammars and dictionaries. Thus, an investigation on constrained or rare combinations of words and *de* (and its variants *da, do*, etc) can be used as basis for identifying MWEs, indicating a preference for frequent combinations such as *greve de fome* (literally *strike of hunger* (hunger strike))

while dispreferring grammatically valid but not adopted variations like *greve da fome* (literally strike of-the hunger).

2 An Overview of *de*

The preposition *de* is a typical linking element in Portuguese, which marks a relation between two or more words. This relation can be of many types, and more or less transparent in its meaning, e.g. the transparent *escada de ferro* (*iron staircase* staircase made of iron) vs the more opaque *pé-de-moleque* (literally *foot of lad*, a candy bar).

Although Rocha Lima [5] describes *de* as a weak preposition, its recognition as expressing relations (of place, origin, etc) suggests that *de* is not as devoid of meaning as initially indicated, e.g. *vir de casa* (come from home) vs *vir para casa* (come home). There is a strong emphasis on viewing *de* as an element on the subcategorization frames of verbs, with the related entries usually appearing as the first ones for *de*. For instance, Moura Neves [6] also begins with *de* introducing complements, in the subcategorization frame of verbs, adjectives, nouns and adverbs. Only then its functions outside the transitivity domain are presented, establishing semantic relations e.g. as adjunct in the VP (*sair de fininho* (leave quietly)) or in the NP (*bolsa de mulher* (ladies' handbag)), and appearing in MWEs like *dar de ombros* (to disregard).

In a corpus-based investigation of MWEs in Brazilian Portuguese, 660 cases (the second highest frequency) are verbs with fixed PP complements [7] and in 85 of them (12.88%) have *de*, like *viver de sonho* (live from dreams). If the contracted forms (*da, das, do* or *dos*) are also considered, there are 150 cases (22.73%), comparing with 94 (14.24%) with the preposition *a* (to) and its variants.

Dictionaries also focus on prepositions linking nouns: (a) Dicionário Aurélio (electronic version of 2001) [8] has 19 (out of 34) entries for *de* linking nouns, while (b) Dicionário Houaiss [9] has 17 (out of 30) entries for *de* linking nouns. On the other hand, the NILC lexicon (available from http://www.nilc.icmc.usp.br/nilc) lists 1,762 (out of 49,551) entries for complement taking nouns with *de* as possible preposition, where 208 entries accept only *de* (*abdicação de* (abdication of)). Among the verbal entries 1,194 (out of 1,418,174) specify *de* in their subcategorization frames, and for 362 of these *de* is the only preposition listed (*abjurar de* (abjurate)). Comparing the entries for *de* in grammars and dictionaries the former emphasize the role of *de* as part of a verbal argument, while the latter have it first as a link between nouns. Moreover, there is no frequency information in them.

3 But just how frequent is *de*?

Its high frequency can be confirmed in several general domain corpora. The 10 most frequent words of the Banco *de* Português (BP)[3], a 240 million word

[3] http://www2.lael.pucsp.br/corpora/bp/index.htm

corpus[4], are shown in table 1, where *de* alone is responsible for 4.42% of the occurrences, while its contracted forms (*da* and *do*) add another 3.32%.

Rank	Word	Freq.	%	Rank	Word	Freq.	%
1	de	1,537,460	4.42%	6	do	609,521	1.75%
2	a	1,082,233	3.11%	7	da	545,271	1.57%
3	o	1,026,380	2.95%	8	em	443,567	1.28%
4	e	726,548	2.09%	9	para	353,847	1.02%
5	que	667,850	1.92%	10	no	308,932	0.89%

Table 1. Most Frequent Words in the Banco do Português

	Folha Corpus			NILC Corpus		Atkins Corpus		Pilla Corpus	
Rank	Word	Freq.	%	Word	Freq.	Word	Freq.	Word	Freq.
1	de	10,491,460	4.70%	de	1,701,990	a	2,008	de	3,013
2	a	7,444,270	3.33%	a	1,243,051	de	1,731	a	2,321
3	o	7,219,765	3.23%	o	1,132,122	o	863	o	1,298
4	e	4,894,555	2.19%	e	838,584	e	777	e	1,288
5	que	4,767,171	2.13%	que	782,252	que	767	se	1,119

Table 2. Most Frequent Words in Corpora

The same pattern is found in both general corpora like the 223 million word Folha Corpus and in the 40 million word segment of the NILC corpus (with 340,016 types), table 2, and also in specialized corpora, like the 1 million word TEXTQUIM Corpus [5] with Chemistry texts[6].

This pattern seems to change, however, in translated texts. We compared two corpora on the same subject, the 34,853 word Atkins Corpus [10], translated from English, and the 54,625 word Pilla Corpus [11], originally written in Portuguese. In the former *de* is not the most frequent word, table 2, and the difference in frequency is statistically significant (p-value of 0.00069 for Fisher's Exact Test 2-tail). This difference in pattern in a translated text may be due to the translation, which seems to have kept the profile and style of the original language.

Analysing in more details the occurrences of *de* in a 1.1 million word subset of the BP corpus, 16,000 contexts for *de* were found. From these, to avoid noise, we kept 2,603 clusters that occur with a frequency of 6 or higher in the corpus.

[4] In the 2003 version.

[5] http://www.ufrgs.br/textquim

[6] In a section of TEXTQUIM containing chapters of chemistry manuals *de* accounts for 5.12% of the 368,005 tokens (against *a* with 3.88%), while in another with Química Nova articles for another 5.45% of the 127,871 words.

Firstly, there is a higher frequency of *de* preceded by nouns than preceded by verbs: in the first 1,000 clusters *de* has a verbal antecedent in only 7.3% of the cases (e.g. *lembrar de*), and in many of them *de* is part of an MWE which may or may not include the verb (e.g. *trabalhar de dia* (work during the day)). In the vast majority of cases *de* is a link between nouns (e.g. *gerente de produção* (production manager)), contrary to the emphasis given to *de* as introducing verbal arguments in the grammars previously discussed. This same tendency is found in specialized corpora from different domains, as shown in table 3 for Chemistry, with the 1,078,580 word TEXTQUIM corpus (section Química Nova), Cookery, with a subsection of the 1,392,706 word CORTEC corpus[7], and Informatics, with the 1,287,260 word POSSAMAI Corpus [12] and the Informatics section of the CORTEC Corpus.[8]

Corpus	# words	Freq. (*de*)	Verbs in top 1000 *de*
BP	199,285	52,326	70
BP (newspaper section)	127,871	9,787	33
TEXTQUIM	127,871	6,980	20
Manual *de* Físico-Química Pilla	54,625	3,013	-
Cookery CORTEC	252,149	22,093	6
POSSAMAI	1,287,260	16,000	12
Informatics CORTEC	207,358	8,434	30

Table 3. *De* in domain specific corpora

These results consistently show the low frequency of *de* with verbs, across different domains. Apart from the verbal contexts the preposition was also found in the following contexts, presented in decreasing order of frequency:

- deverbal nominal constructions: *elevação de temperatura* (increase in temperature);
- terms and multiword expressions: *taxa de transmissão* (transfer rate);
- conventional phrases: *ponto de partida* (starting point).

The degree of specialization seems to have some influence on this frequency, as indicated by the statistically significant difference in frequency of *de* with verbs for the same domain: 0.075% for POSSAMAI and 0.35% for Informatics CORTEC (Fisher's Exact Test with a 2-tail p-value of 0.0000016). This suggests that the more general the domain is the higher the frequency of *de* with a verb is. Moreover, these results with *de* linking nouns more often than verbs suggest that the use of this preposition in corpora is more compatible with that recorded in the dictionaries than with that in the grammars.

[7] http://www.ffich.usp.br/dlm/comet

[8] The two Informatics corpora differ in their degree of generality: the POSSAMAI Corpus is more specialized and contains conference papers, while the CORTEC Corpus is more general with articles from newspapers and from the INFOEXAME magazine.

3.1 Lexical Preferences

In BP corpus 6 out of the 10 more frequent clusters have a noun immediately preceding *de* (*gerente de* (manager of - 168 occurrences)), and the first verb not in an MWE is ranked 92 in frequency (*sei de* (I know of)). In the WebCorp (http://www.webcorp.org.uk/) with Google as search engine 3,110 concordances were generated, with *de* linking nouns in most of them, and the first case of a verb preceding the preposition appears only after the 150th context.

Even if there is a large number entries for verbs combining with *de* in resources (in the NILC lexicon, 1,194 entries are listed as combining with *de* against 1,058 entries with a), an analysis of the verbs found in the first 1,000 clusters of the BP corpus shows only 38 verbs in 73 instances. This suggests that even though a large number of verbs can be combined with this preposition, in real data, only a few of these actually occur, and among them the most frequent are *ser, gostar* and *deixar*, with 11, 6 and 4 occurrences respectively. Moreover,among BP's 50 most frequent clusters with verb+*de* there were 23 verb types in 510 sentences. After removing expressions (*a partir de* (from)) and merging inflected forms of verbs, 19 verb types remained with 418 occurrences. The most frequent of these are *ser* (be - 63 occurrences), *andar* (walk - 47), and *vir* (come - 40), and their frequencies are much lower than those of the top ranked nouns (cited above). In terms of syntax, most cases are oblique verbs with *de* introducing the oblique complement (e.g. *gostar de* (like)), with 10 types and 215 occurrences; 3 circumstantial verb forms (*ser, andar* and *partir* (be, walk and leave) - 110 cases), where *de* further specifies the semantics of the verb (e.g. with information about origin); and 3 verbs where *de* is an adjunct (e.g. *aferir* (measure) - 33 cases). All of these 23 verbs are very general and frequently used words, already listed in the NILC lexicon.

4 Conclusions and Future Work

This study provided a first step towards understanding the characteristics of the preposition *de*, the most frequent word in use in Brazilian Portuguese. Through the analysis of the information in lexical resources, and a corpus-based investigation, the contexts in which *de* occurs were discussed. The results obtained show *de* primarily as a nominal link rather than a verbal one. Moreover, in corpus data only a small number of verbs preceding *de* were found, despite its enormous potential for combination (it is the most frequently specified preposition in verbs' PP arguments in the NILC lexicon). Further investigation into the proportion of *de*-taking verbs in relation to all verbs in corpora is planned for the future. Adjuncts were not the most common function of this preposition, as it occurred much more often as part of the oblique complement of a verb. The results described in this paper can be used to inform NLP efforts. For instance, as all of these verbs found with *de* in corpora are frequently used and general words, which will probably be contained in a lexical resource, an NLP system would possibly already be equipped to deal appropriately with cases of *de* in verbal contexts.

This investigation can also be used as the basis for larger scale studies on the automatic identification of MWEs containing *de*, for applications like Machine Translation where an adequate treatment can lead to more natural results. As a next step, we plan to extend these results looking at characteristics of verbs and nouns used with this preposition. Investigation will also proceed on the use of lexical profiles as indicators of the quality/naturalness of a translations.

Acknowledgements

We would like to thank projects FAPERGS/BIC-2008 08513752, and CNPq PQ301102/2006-6 for providing the necessary support for this work.

References

1. Bame, K.: Aspectual and resultative verb-particle constructions with up. In: Ohio State University Linguistics Graduate Student Colloquium. (1999)
2. Abeillé, A., Bonami, O., Godard, D., Tseng, J.: The syntax of french à and de: an hpsg analysis. In Saint-Dizier, P., ed.: Linguistic Dimensions of Prepositions. Kluwer (2006)
3. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002). Volume 2276 of (Lecture Notes in Computer Science)., London, UK, Springer-Verlag (2002) 1–15
4. Dias, G., Lopes, L.: Extração automática de unidades polilexicais para o portugues. In Sardinha, T.B., ed.: A Lingua Portuguesa no Computador. Mercado de Letras (2005)
5. Rocha Lima, C.: Gramática Normativa da Língua Portuguesa. José Olympio, Rio de Janeiro (1980)
6. Neves, M.: Gramática de Usos do Portugues. UNESP, São Paulo (2003)
7. Vale, O.A.: Expressões Cristalizadas do Portugues do Brasil: uma Proposta de Tipologia. PhD thesis, Unesp, Araraquara (2001)
8. Ferreira, A.: Dicionário Aurélio Eletronico. Novo Dicionário Aurélio - Século XXI. Nova Fronteira/Lexikon Informática, São Paulo (1999)
9. Houaiss, I.A.: Dicionário Eletronico Houaiss da Lingua Portuguesa. Editora Objetiva, Rio de Janeiro (2001)
10. Atkins, P.: Físico-Química. 6 edn. Volume 1. Livros Técnicos e Científicos (1998)
11. Pilla, L.: Físico-Química. 1 edn. Volume 1. Livros Técnicos e Científicos (1979)
12. Possamai, V.: Marcadores textuais do artigo científico em comparação portugues-ingles um estudo sob a perspectiva da tradução. Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre (2004)