DEVISING A SKETCH GRAMMAR FOR ACADEMIC PORTUGUESE

Tanara ZINGANO KUHN
Faculty of Letters, University of Lisbon
Centre of General and Applied Linguistics Studies (CELGA-ILTEC), University of Coimbra

Iztok KOSEM
Faculty of Arts, University of Ljubljana &
Trojina, Institute for Applied Slovene Studies

Kuhn, T. Z. and Kosem, I. (2016): Devising a sketch grammar for academic Portuguese. Slovenščina 2.0, 4 (1): 124–161.
DOI: http://dx.doi.org/10.4312/slo2.0.2016.1.124-161.

This paper presents the development of a new sketch grammar designed specifically for CoPEP, a newly compiled 40-million corpus comprising texts from academic journals, tagged with Freeling v3, the default tagger available in the Sketch Engine for corpora of Portuguese. We first provide an overview and evaluation of existing sketch grammars for Portuguese, followed by a detailed description of the development of a new sketch grammar, and the presentation of some of the problems encountered. We conclude by summarizing the main findings, highlighting important implications, and offering suggestions for further improvement of the sketch grammar. More accurate and varied word sketch results than those offered by the current default sketch grammar indicate that our sketch grammar can be used for advanced lexicographic tasks such as automatic extraction of lexical data from CoPEP, the methodology of knowledge acquisition planned for the compilation of the proposed dictionary of Portuguese for university students. Moreover, this new sketch grammar can be used with any other corpus of Portuguese tagged with Freeling v3, which makes it an important resource for lexicographic and corpus linguistic research of the Portuguese language.

Keywords: sketch grammar, Portuguese, corpus, dictionary, evaluation

1 INTRODUCTION

State-of-the-art lexicographic methods have moved on from an analysis of only concordances and lists of collocations of a word. In the last decade, the combination of collocation and grammar, offered through the functions such as Word sketch in the Sketch Engine corpus tool (Kilgarriff et al. 2004), has been used more and more frequently. In fact, as Atkins and Rundell (2008: 110-111) suggest, word sketches have even become a departure point for lexicographic analysis. Moreover, word sketches are also at the core of the semi-automated approach to dictionary compilation, as originally proposed by Rundell and Kilgarriff (2011) and first implemented into lexicographic practice in Slovenia (see Kosem et al. 2013; Logar and Kosem 2013; Kosem et al. 2014; Gantar et al. 2016) and Estonia (Kallas et al. 2015).

In order to build word sketches, two conditions have to be met. One is a POS-tagged corpus,[1] and the other is sketch grammar, i.e. definitions of grammatical relations for the language, using corpus query language (CQL). For several languages, sketch grammars already exist, but an evaluation is needed to determine their suitability for the purposes of a particular lexicographic project. This was also the case with our project, namely designing a corpus driven dictionary of Portuguese for university students (*Dicionário de Português para estudantes universitários*, hereafter DOPU), where in the end not only a new sketch grammar was needed, but also a new corpus had to be compiled due to unsuitability or unavailability of existing corpora of academic Portuguese.

This paper presents the development of a new sketch grammar designed specifically for academic Portuguese. In Section 2, we provide an overview of existing corpora containing academic Portuguese, and point out their shortcomings in terms of corpus-based research of academic language,

[1] The other option is using a parsed corpus, but as Kilgarriff and Kosem (2012) point out, those corpora are rarely available.

followed by the description of the corpus compiled for our project and used in the development of the new sketch grammar. Section 3 is dedicated to the new sketch grammar, and offers an overview and evaluation of existing sketch grammars for Portuguese, a detailed description of the development of the sketch grammar, and the presentation of some of the main problems encountered. In Section 4, we summarize the main findings, highlight important implications of our research, and provide suggestions for further improvement of the sketch grammar.

2 CORPORA OF ACADEMIC PORTUGUESE

DOPU's target users are students in higher education, attending courses in different areas of knowledge, whose language of instruction is (Brazilian or European) Portuguese and thus need to read and write academic texts in Portuguese. As a corpus driven dictionary it must portray the linguistic information that is based on texts that reflect the way language is used by expert writers from Brazil and Portugal in academic settings in different areas of knowledge. Hence, the corpus needed for making DOPU must have the following characteristics:

- composed of academic written texts portraying exemplary language
- balanced in terms of Portuguese varieties: 50% of Brazilian Portuguese, 50% of European Portuguese
- covering different academic areas
- synchronic
- large in size.

The first step was to examine existing Portuguese corpora containing academic texts and determine their suitability for our research. Out of many corpora of Portuguese in existence,[2] which cover different language varieties, registers,

[2] For collections of available corpora, see http://www.linguateca.pt/, http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources, http://clul.ul.pt/en/resources.

and genres, only few comprise academic texts. As Table 1 shows, although existing corpora of Portuguese do contain academic texts, none of them fulfils all our criteria: balance between varieties, in terms of areas of knowledge and words per area; academic written texts portraying exemplary language; synchronic; large size; and detailed metadata. Consequently, a decision was made to compile a new corpus of academic texts, which we named CoPEP.

Corpus and author(s)	Size	Characteristics	Reasons for not suiting our purposes
Portuguese Web 2011 (ptTenTen, Palavras parsed) Authors: The Sketch Engine team	2,757,635,105 words[3]	Texts from sites of academic/scientific nature (universities, journals, governmental, thesis repositories, etc.) Parsed by PALAVRAS dependency parser (Bick 2000).	Crucial metadata such as source (type of publication: journal, book, thesis, etc.), year of publication and area of knowledge are not available. No possibility to measure quality of writing and corpus composition.
Portuguese Web 2011 (ptTenTen, Freeling v3) Authors: The Sketch Engine team	3,900,501,097 words	Texts from sites with academic/scientific nature (universities, journals, governmental, thesis repositories, etc.) Tagged by Freeling 3.0 (Padró and Stanilovsky 2012)	Crucial metadata such as source (type of publication), year of publication, area of knowledge and language variety are not available. Country of the website is made equivalent to language variety, which is not an accurate approach for determining such relevant information. No possibility to measure quality of writing and corpus composition.
Corpus Araneum Portugallicum Maius (Portuguese, 15.05) 1,20 G as a language resource Author: Vladimír Benko	862,134,902 words	Texts from sites of academic/scientific nature (universities, journals, governmental, thesis repositories, etc.). To be used for contrastive linguistics and bilingual lexicographic projects.	Crucial metadata such as source (type of publication: journal, book, thesis, etc.), year of publication, area of knowledge and language variety are not available. No possibility to measure quality of writing and corpus composition.
Corpus Brasileiro (the Brazilian Corpus) Author: Tony Berber Sardinha (coordinator)	1,133,416,757 tokens	General corpus of Brazilian Portuguese. Academic subcorpus contains 258,585,002 tokens from articles, 310,972,387 tokens from theses and dissertations, and 6,947,244 tokens	Crucial metadata such as year of publication and area of knowledge are not available. No information on quality of texts comprising the academic subcorpus. Only Brazilian Portuguese.

[3] In this paper, we use words or tokens, or both, when providing the information on corpus size, depending on the information that is available.

		from annals.	
Corpus do Português (Genre/historical version) (the Corpus of Portuguese) Authors: Mark Davies and Michael Ferreira	45 million words	Texts of the 1300s to the 1900s. The texts from the 1900s make up 20 million words, with balance between academic, fiction, spoken and newspaper genres. Its academic subcorpus consists of 3,087,052 words from Portugal and 2,816,802 from Brazil.	Academic subcorpus is composed of entries retrieved from Brazilian and Portuguese online encyclopaedias.
CPBA – *Corpus do Português Brasileiro Acadêmico* (the Academic Brazilian Portuguese Corpus) Authors: The research group UPLA, coordinated by Cristina Becker Lopes Perna, at PUCRS (Brazil)	22,777,993 tokens (Peixoto 2015: 44)	Books and journals from six different areas of knowledge provided by eight Brazilian universities comprising written productions of professors and (undergraduate and graduate) students.	Not publicly available. Only Brazilian Portuguese.
CRPC - *Corpus de Referência do Português Contemporâneo* (Reference Corpus of Contemporary Portuguese) Authors: developed at the *Centro de Linguística da Universidade de Lisboa* (CLUL).	311 million words (spoken+written) approx. 310 million words of written texts	General language corpus. European Portuguese and other varieties (Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, S. Tome and Principe, Goa, Macao and East-Timor). Comprising different text types, including scientific. Texts from the second half of the 19th century to 2008.	Metadata not consistently available.

Table 1: Summary of existing corpora of Portuguese containing academic texts.

CoPEP - *Corpus de Português Escrito em Periódicos* (Corpus of Written Portuguese in Academic Journals) (Kuhn and Ferreira 2016) is composed of texts published mainly between 2000 and 2016 (2% of texts are from the 1990s) in online academic journals that are part of SciELO (Scientific Electronic Library Online), an open access platform containing collections of journals from Brazil (SciELO-Br) and Portugal (SciELO-Pt). Besides not having access restrictions, SciELO gathers journals from different areas in a single website, which facilitates text crawling. In addition, the delicate issue of journals allocation into domains is avoided due to the common organisational structure adopted by all collections, which follows Capes classification[4] of areas of knowledge in Great Areas.

Furthermore, SciELO's strict criteria for admission and retention of journals in its collections, such as, among others, scientific content, peer-review process, journal usage and impact factor, imply texts of high quality (in both content and language) from different domains. For the corpus, this means SciELO-Br and SciELO-Pt provide samples of exemplary academic Portuguese writing. Due to the need of balance between the two Portuguese varieties, corpus size was determined by the size of the smallest text collection, SciELO-Pt. The CoPEP corpus contains 9.859 texts, distributed among six Great Areas grouped in three Schools of knowledge, totalling 40,246,492 words.

Table 2 presents the statistical information on the corpus contents. As we can see, the two subcorpora, comprising texts written in Brazilian Portuguese and European Portuguese respectively, are nearly the same size. Similar balance in size of both varieties can be also found for Great Areas and Schools, whereas the balance between Great Areas and Schools is very much in favour of Humanities. This imbalance is the reflection of the distribution of published

[4] Capes stands for *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (Coordination for the Improvement of Higher Education Personnel) and is a foundation within the Ministry of Education in Brazil. It has created the Table of the Areas of Knowledge of Higher Education, with four hierarchical levels, going from the most general – Great area – to the most specific – Speciality.

documents per Great Area on the SciELO platform.

	Whole corpus	Brazilian Portuguese	European Portuguese
Texts	9,859	3,795	6,064
Words	40,246,492	20,149,980	20,096,511
Tokens (also for data below)	48,548,527	24,306,513	24,242,014
Humanities	**30,955,740**	**15,458,157**	**15,497,583**
Human Sciences	25,570,856	12,763,567	12,807,289
Applied Social Sciences	5,384,884	2,694,590	2,690,294
Life Sciences	**16,118,303**	**8,099,937**	**8,018,366**
Health Sciences	13,515,763	6,787,116	6,728,647
Agricultural Sciences	2,602,540	1,312,821	1,289,719
Exact, Technological and Multidisciplinary Sciences	**1,474,484**	**748,419**	**726,065**
Exact-Earth Sciences	793,877	400,040	393,837
Engineering	680,607	348,379	332,228

Table 2: Statistical information on CoPEP.

Although CoPEP is relatively small according to modern day corpus standards, its size makes it ideal for evaluation and tool development, which are usually done on sample corpora with sizes varying from 50 million to 100 million tokens, or sometimes even smaller than that. The corpus was uploaded into the

Sketch Engine, where it was tokenised, lemmatised and tagged with the Freeling 3.0 tagger (Padró and Stanilovsky 2012).

3 SKETCH GRAMMARS FOR PORTUGUESE

Sketch grammar is a file with grammatical relations, or gramrels, and processing directives for the Sketch Engine system to compute different types of relations through statistics calculations. The data obtained with these computations then form the basis of the Word Sketch feature in the Sketch Engine, and relatedly, the Thesaurus and Sketch Diff features.

Sketch grammars devised for POS-tagged corpora use regular expressions over POS-tags to find matches for grammatical relations. Queries are written in Corpus Query Language (CQL), with attribute-values names following the tagset used for corpus tagging.

There are currently four different sketch grammars for Portuguese available in the Sketch Engine. The first one is the Sketch Grammar for Portuguese, FreeLing tagset version 1.1[5] (henceforth FreelingSkG), the Sketch Engine's default sketch grammar for all corpora of Portuguese tagged with the Freeling tagger. In fact, this is the sketch grammar devised for the Spanish TenTen corpus, which uses the same tagger, meaning that the queries were originally meant for capturing grammatical relations in Spanish, not Portuguese.

The second sketch grammar for Portuguese is the one used for the PtTenTen [2011, Palavras parsed] corpus[6] (henceforth PalavrasSkG)[7] and devised especially for the compilation of the Oxford Portuguese Dictionary (2015). The Sketch Engine computes word sketches automatically from these kinds of

[5] This is the version we used in our research. The latest version is 1.1.1: https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/pttenten11_freeling_v3_1.
[6] This corpus has been renamed to Portuguese Web 2011 (ptTenTen, Palavras parsed).
[7] Although this is still the sketch grammar portrayed when users click on gramrels names or Word Sketch in Corpus Info, recently the actual word sketches computed and gramrels names have changed.

parsers, so PalavrasSkG does not contain CQL-written queries, but only the names of the grammatical relations.

Two other sketch grammars for Portuguese were available at the time. The first one is the Compatible Portuguese Sketch Grammar definition (henceforth AraneaSkG) (Benko, 2014b), devised for the Aranea (Web) Corpora Family (Benko 2014a). The purpose of AraneaSkG is to provide uniform word sketch results among different languages, i.e. it is intended to be language non-specific. Consequently, sketch grammars are not syntactically-based like other sketch grammars available in the Sketch Engine, but rather collocationally-based. The second one is the Portuguese word sketches (Linguateca parsed data) version 1 (henceforth LinguatecaSkG), written for the Cetenfolha, Cetempublico corpus,[8] a 56-million-word corpus comprising extracts of texts published in Folha de São Paulo, a Brazilian newspaper, and in Público, a newspaper from Portugal. Although it was processed by PALAVRAS dependency parser (Bick 2000), LinguatecaSkG contains CQL-written queries where both regular expressions over POS-tags pattern matching and Constraint Grammar tags are used.

Preliminary tests of the performance of the default sketch grammar for Freeling-tagged corpora in the Sketch Engine using a sample corpus of 5 million words (henceforth the sample-5mil corpus), comprising files from SciELO-Pt, revealed several problems, such as: in majority of cases, words with capital letters were tagged as proper nouns, independently of their actual category; word sketches of a number of lemmas indicated that participial adjectives were never matched in gramrels with adjectives – we soon discovered that participle forms were tagged as verbs by Freeling 3.0; some word sketches returned empty results, in some cases word sketches yielded wrong results.

Such poor performance led us to the conclusion that the default sketch

[8] This corpus has been renamed to Newspapers in Portuguese (CetemPúblico, CetenFolha) https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/portuguese.

grammar could not be directly applied to CoPEP, hence a new sketch grammar had to be developed for our corpus. The first step of this process was to evaluate, besides the default grammar, other existing sketch grammars for Portuguese, in order to determine whether they, or their parts, could be used for our purposes.

3.1 Evaluation of FreelingSkG and PalavrasSkG

It was necessary to compare FreelingSkG with another sketch grammar, so that there would be standards for deciding which gramrels could be maintained in full, which ones would need to be revised, and identify any missing gramrels for which completely new queries would be required. PalavrasSkG was chosen to be the contrasting sketch grammar because it had been devised especially for the compilation of a dictionary of Brazilian and European Portuguese varieties. AraneaSkG and LinguatecaSkG were used at a later stage, when developing the new sketch grammar; the queries from the two sketch grammars were compared to the ones being developed in order to provide input or better alternatives.

The evaluation focussed on the coverage and accuracy of queries for different gramrels. We used a sample of lemmas (adjectives, adverbs, verbs, and nouns) from the sample-5mil corpus. We examined separately gramrels for each word class in word sketches of sample lemmas, based on the two corpora (the ptTenTen11 [2011, Freeling v3] corpus[9], and the ptTenTen [2011, Palavras parsed] corpus). The results for all the identified gramrels were evaluated by analysing samples (up to 250 concordances) of three collocates (one from the top, one from the middle, and one from the bottom of the list of the first 25 collocates ordered by salience) in order to verify if the results were valid for the gramrel in question, and to investigate clues indicating which queries were evaluated (the latter for PalavrasSkG only); in addition, notes of any errors were taken. Finally, possible missing grammatical relations for the selected word

[9] This corpus has been renamed to Portuguese Web 2011 (ptTenTen, Freeling v3).

class were noted, and potential queries for those relations were recorded.

3.1.1 EVALUATION OF FREELINGSKG

With regards to FreelingSkG, certain errors were envisaged due to the fact that the sketch grammar was originally written for Spanish rather than Portuguese. One example is the symmetric gramrel `=and_or`, which returns wrong results. This is due to the use of words *y* and *o* in the gramrel, which are the Spanish equivalents of the English words 'and' and 'or', respectively. For Portuguese, the words *e* and *ou* should be used.

Another significant source of errors were additional issues regarding corpus tagging, besides the two previously mentioned (tagging capitalised words as proper nouns and participial adjectives as verbs). Tagging errors were also the cause of empty and wrong results.

Evaluation revealed that gramrels `=adj_complement` and `=predicate`, whose queries display semiauxilary verb *ser* ('to be') (tag =VS) in position 1, were never displayed in the output panel of word sketches of the selected lemmas. In order to examine the problem further, CQL searches of the queries were performed on the corpus, always returning empty results. A detailed analysis revealed that, although in the tagset VS stands for semiauxiliary verb (verb 'to be'), there are no tokens annotated with that tag in the sample corpus (nor in CoPEP). The verb *ser* is tagged with VM, i.e. as a main verb. This explained why the word sketch output for those gramrels was empty.

Wrong results due to tagging errors were spotted when the examination of word sketch output of several lemmas showed discrepancy between the word class of the collocate(s) and the word class defined in the gramrel. For example, it was unexpected to see verb+noun collocations such as *apresentar olhar* and *apresentar caráter* displayed in the table of results of the gramrel `=object_inf` (verb as keyword and verb-infinitive as collocate) for the keyword *apresentar* ('to present/show'). A close examination of the tags of the collocates revealed

these nouns had been tagged as verbs. That indicates that the program probably interpreted the -ar and -er endings in words such as *olhar* ('look') and *caráter* ('character') as markers of the infinitive form of verbs belonging to first and second conjugation respectively. Many other cases of inconsistency between gramrels and their results were found and different types of tagging errors recorded.

In addition to the accuracy of FreelingSkG being compromised due to Spanish framed definitions and tagging errors, examination showed that FreelingSkG is also rather limited in terms of query coverage. Some noticeable examples are: no gramrels for adverbs as keywords; no gramrel for the pair adjective-noun when adjective is prenominal; gramrels with adjectives do not include participial adjectives.

Given these findings, we concluded that FreelingSkG could not be employed for the analysis of CoPEP without considerable improvement to its queries. The evaluation of word sketches of sample lemmas has also indicated that attention should be paid to the tagset and tagging issues when preparing the sketch grammar for CoPEP, or in fact any corpora tagged with Freeling.

3.1.2 EVALUATION OF PALAVRASSKG

The evaluation of PalavrasSkG was conducted on the ptTenTen [2011, Palavras parsed] corpus, which is dependency parsed by PALAVRAS (Bick 2000). Since the Sketch Engine computes word sketches automatically from the parsing output, the sketch grammar for this corpus is composed of a list of gramrels without queries. As a result, the evaluation also involved investigating clues to components and structures of queries for different gramrels.

The analysis of the coverage of gramrels and of the accuracy of queries has revealed a number of interesting points, such as:

a) Dependency relations (deprels) annotation allows capture of collocations with a very large span window between a keyword and a

collocate. For instance, the collocation *paciente apresentar* ('patient', noun ; 'to show', verb) for the gramrel `=N subj_of %w_V` was captured in a sentence despite a 15-token-long relative clause between the keyword and the collocate.

b) Deprels annotation allows matching of inverted constructions. For the case of personal verbal passive constructions, i.e. agent of the passive + verb 'to be' + main verb in the participle form, simple position-based queries, which follow the canonical subject + verb + object order, detect the first element as a subject of the main verb, while it is, in fact, its object. Thus, with deprels annotation, verb-object collocations are captured, regardless of the keyword and noun positions in the sentence.

c) Although there were only 13 gramrels in PalavrasSkG, deprel attribute view option showed annotation of many more relations than stated. For instance, when analysing the word sketch output for the verb-object pair relations, with verb as the keyword, some other deprels involved in matching such a collocation were `# V comp %w_V` or `# ADJ _por_ %w_V`, among others.

d) Identification of tagging errors when adjectives were in prenominal position. Such constructions are marked in Portuguese and were not explicitly covered by PalavrasSkG. Thus, CQL searches of sample lemmas (adjectives) followed by a noun were performed to confirm this missing gramrel coverage. To our surprise, concordances seemed to contain good collocations, i.e. the sample adjective lemmas followed by correct noun collocates, but a detailed inspection showed that those were false positives, since the original collocation adjective+noun was only matched due to wrong tagging of both the keywords and the collocates. In other words, adjective lemmas had been tagged as nouns and noun lemmas as adjectives.

e) Identification of bad collocations. Errors in finding collocates seem to be a result of parsing errors. For instance, in sentences with two clauses linked by the conjunction *e* ('and'), where the subject of the first clause was matched with the main verb of the second clause.

Although PalavrasSkG could not be applied to CoPEP due to its parsing (and not POS-tagging) annotation, the findings of this evaluation have had significant implications for the understanding of this sketch grammar. First, several possible grammatical relations in Portuguese were recorded, with special attention paid to categories and occurrences of items between keywords and collocates. Second, dependency relation annotation proved to be particularly useful for finding relations between items that have several intermediary words, and in inverted constructions. Finally, occasional failure in capturing collocations helped us record different errors for future reference. Overall, it was found that PalavrasSkG covers an extensive set of grammatical relations and contains very accurate queries.

The evaluation of FreelingSKG and PalavrasSkG has revealed advantages and shortcomings of both sketch grammars, as well as several issues of the Freeling tagger, which has been used for tagging our corpus. Nonetheless, the overall conclusion was that neither of these sketch grammars could be used for our purposes, but rather that a completely new sketch grammar for academic Portuguese would need to be developed; this new sketch grammar could, however, still utilize some of the good gramrel queries, or parts of them, from the above evaluated sketch grammars.

3.2 Devising a new sketch grammar for academic Portuguese

Devising the sketch grammar for academic Portuguese (henceforth AcadPortSkG) consisted of two phases: writing gramrel queries and evaluation. The first phase was grounded in a trial and error method, where queries were written and tested many times until satisfactory results were reached. This process was not only laborious but also time consuming: for every new attempt,

the corpus had to be recompiled in the Sketch Engine. To speed up corpus recompilation and analysis, a sample 1-million-word corpus was used instead of the entire corpus. Once the results of the sketch grammar were deemed satisfactory enough, we proceeded to the second phase, which was conducted on the entire CoPEP, recompiled with the new sketch grammar.[10]

All this work came down to a sketch grammar with symmetric (1), unary (3), dual (14), and trinary (2) grammatical relations covering attributive (pre- and postpositional) and predicative adjectives; nouns as predicative complement, subjects, and objects of verbs (unmarked order); prepositional phrases with nouns and verbs; infinitive as verb/noun/adjective complement; impersonal and personal verbal passive constructions; impersonal constructions with *se*; verbs followed by *que*-clauses (subordinate clauses); verbs with gerund as a complement, and adverb-verb and adverb-adjective pairs. An example of the word sketch output is shown in Figure 1.

estudo *(noun)*
CoPEP freq = 81,688 (1,682.60 per million)

sintagma preposicional	70,299	
...deestudo	22,327	0.27
...emestudo	15,212	0.19
estudo de N	14,147	0.17
estudo sobre N	3,627	0.04
estudo em N	3,601	0.04
...paraestudo	2,132	0.03
estudo com N	1,861	0.02
...aestudo	1,826	0.02
...porestudo	1,289	0.02
...comestudo	1,045	0.01
estudo por N	1,018	0.01

estudo_N suj de V	32,156	0.39
mostrar +	1,030	9.42
demonstrar +	809	9.36
ir +	4,092	9.19
estudo foi		
ter +	2,518	9.01
revelar +	648	8.85
apontar +	498	8.71
estudos apontam		
avaliar +	410	8.57
permitir +	610	8.38
utilizar +	386	8.33

estudo_N mod por Adj-Part	19,779	0.24
realizar +	2,262	11.27
estudo realizado		
recente +	771	9.98
estudo recente		
anterior +	666	9.57
estudos anteriores		
empírico +	540	9.48
estudos empíricos		
comparativo +	355	9.07
estudo comparativo		
publicar +	364	9.03

Figure 1: Partial word sketch for *estudo* ('study', noun).

The results of trinary relations open on a separate page, allowing detailed

[10] The current version is 9. At the time of the experiment, version 7.1 was used. For every new version of AcadPortSkG, the results of changes were evaluated in the same manner as here described.

analysis of each relation. For instance, there are 35 gramrels for *estudo* followed or preceded by a prepositional phrase (the gramrel column titled *sintagma preposicional*), each of them with their own column of collocates. So for example, of the 22,327 occurrences of **...de estudo**, the collocation *resultado do estudo* ('result of study') represents 1,521 occurences; similarly, of 14,147 occurrences of **estudo de N**, the collocation *estudo de caso* ('case study') represents 1,317 occurences.

To give some indication of the number of gramrels per lemma, we provide the data for the top five frequent lemmas per word class (Table 3). We can see that the generalisation of number of gramrels per word class can only be made to some extent. It is possible to affirm which gramrels do not take certain word classes as keywords, namely, the three unary relations: no adverbs and nouns; both types of trinary relations: no adverbs; and prepositional phrase trinary relations: no adjectives. Besides that, numbers vary according to the characteristics of the keyword in question.

		Symmetric	Unary	Dual	Trinary	total
Top 5 nouns	estudo	1		10	Prep.phrase: 36 Prep+inf: 7	54
	relação	1		10	Prep.phrase: 36 Prep+inf: 6	53
	forma	1		10	Prep.phrase: 35 Prep+inf: 7	53
	ano	1		10	Prep.phrase: 31 Prep+inf: 9	51
	trabalho	1		10	Prep.phrase: 37 Prep+inf: 8	56
Top 5 verbs	ser	1	3	7	Prep.phrase: 23 Prep+inf: 12	46
	ir	1	3	8	Prep.phrase: 24 Prep+inf: 11	47
	ter	1	3	11	Prep.phrase: 22	46

					Prep+inf: 9	
	poder	1	2	6	Prep.phrase: 17 Prep+inf: 9	35
	estar	1	3	8	Prep.phrase: 22 Prep+inf: 7	41
Top 5 adjectives	social	1	1	5	Prep+inf: 8	15
	maior	1	1	5	Prep+inf: 6	13
	novo	1	1	5	Prep+inf: 7	14
	político	1	1	5	Prep+inf: 7	14
	primeiro	1	1	5	Prep+inf: 4	11
Top 5 adverbs	não	1		4		5
	mais	1		4		5
	também	1		4		5
	assim	1		4		5
	ainda	1		4		5

Table 3: Numbers of gramrels for the five most frequent lemmas in each word class.

3.2.1 PHASE 1: WRITING ACADPORTSKG

The method of writing AcadPortSkG was as follows:

1. Select part-of-speech items (noun, adjective, verb, adverb);
2. Determine grammatical relations between them;
3. Name those gramrels;
4. Define directives to find gramrels;
5. Write queries for gramrel matching in the sample corpus:
    a. Use CQL concordance search in the sample corpus to verify regex matching;
    b. If regex query works, write it in the sketch grammar file;

6. Recompile the sample corpus after each change to the sketch grammar;
7. Verify in the sample corpus if the sketch grammar works.
8. Once the sketch grammar yields satisfactory results, go to phase 2.

As an illustration of what writing a sketch grammar entails, a brief account of the process of writing queries for the grammatical relations between adjective and noun is given below.

As mentioned earlier, preliminary annotation test pointed out that participle forms were tagged as verbs only, although in Portuguese those forms can also function as adjectives. Thus, for gramrels with this category, a tag for the participle form of the verb (V.P.*) had to be added.

Firstly, a simple combination of a noun followed by adjective/verb participle (unmarked order in Portuguese) was tested. The name used for this gramrel was `=N mod por Adj-Part`. As expected, majority of collocations identified for the evaluated lemmas were valid, for example, at the lemma *social*, *ciências sociais* ('social sciences') and *classe social* ('social class').

The marked order of adjectives in Portuguese, i.e. before nouns, is not covered by FreelingSkG. Thus, the directive *DUAL for the two gramrels `=N mod por Adj-Part/ =Adj-Part mod N` was added. Word sketches of test lemmas yielded good matches; for example, for the lemma *estudo* ('study', noun), collocations like *estudo analítico* ('analytical study') for the former gramrel, and *presente estudo* ('present study') for the latter.

After confirming these two gramrels worked fine on the sample 1-million-word corpus, different intervening words were tried out. For that, new queries were written and searched in the sample corpus. If these queries produced good results, they would be included in the test sketch grammar, and after each such change was implemented, the sample corpus had to be recompiled in the Sketch Engine.

To sum up, the experiment involved, firstly, adding one optional adverb, then

one optional adjective following the optional adverb. Analysis of word sketches of a number of lemmas showed that these two extra optional items increased the number of good matches for the gramrel noun+adjective. For the reversed gramrel, i.e. adjective+noun, the second adjective preceding the noun returned bad matches in most cases, whereas adverbs were not tested because they do not occur in this position in Portuguese (cf. Perini 2002). Next, the number of optional intermediary adverbs and adjectives were expanded to two in the gramrel `=%w_N mod por Adj-Part`, which yielded more results (which were still good) than its initial version. Hence, collocates within a wider span were found, for example, the adjective *realizado* was captured for the keyword *estudo* in *Em estudo ainda não publicado realizado*. Here, there is a three-space window comprising two adverbs (*ainda* and *não*) and one adjective (*publicado*).

Finally, an attempt to capture adjectives collocating with the head of a noun phrase composed of head noun + prepositional phrase[11] + adjective led to the inclusion of optional intermediary preposition and determiners. Since the examination of concordances revealed much noise, which severely hindered the accuracy of collocation matching, these items were excluded.

In the end, these were the gramrels devised for capturing collocates between adjective and noun (R stands for adverbs):

```
*DUAL
=Adj-Part mod %w_N/%w_Adj-Part mod N
2:"A.*|V.P.*" 1:"N.*"
*DUAL
=%w_N mod por Adj-Part/N mod por %w_Adj-Part
1:"N.*""R.*"{0,2}"A.*|V.P.*"{0,2}2:"A.*|V.P.*"
```

[11] Perini (2002) uses the term *modifier* to refer to (preposed and postposed) words that modify the head of noun phrases. According to him, "a modifier can also be composed of a prepositional phrase" (ibid.: 327). Prepositional phrases are contiguous to the heads when such a phrase is a classifier; in case of a second modifier (an adjective), this comes at the end of a noun phrase, as in the example above.

It is noteworthy that this phase of writing gramrel queries not only resulted in a new sketch grammar for academic Portuguese, but also contributed to the improvement of the overall quality of our research due to two important revelations, which, in turn, led to correction measures.

Firstly, verification of regex matching through CQL concordance search in the sample corpus revealed 'junk' in the corpus, such as the following textual passages "[Creative_Commons_License]", "texto apenas em PDF", "abstract", "resúmen", besides email addresses, phone numbers, and texts written in languages other than Portuguese. Consequently, extra cleaning of the corpus was performed and its quality significantly enhanced.

Secondly, other types of tagging errors, in addition to the ones already listed in the previous sections, were spotted during this phase. Attempts to work around problems related to the identified tagging errors demanded unique approaches to different gramrels, making the whole process more complex. A few examples of such workarounds are discussed in the next section.

3.2.2 PHASE 2: EVALUATION OF ACADPORTSKG ON THE COPEP CORPUS (40 MILLION WORDS)

After developing AcadPortSkG on a sample corpus, we moved to its evaluation on the CoPEP data. This entailed compiling the corpus in the Sketch Engine using AcadPortSkG, defining a methodology of evaluation, conducting the evaluation, and proposing workarounds for gramrels in which annotation problems seriously affected the results.

The objective of the evaluation was to verify whether the devised gramrel queries captured correct information. A sample lemma list for the evaluation was selected according to the following two criteria: frequency and diversity of characteristics of each word class. By diversity, we mean heterogeneity of word class characteristics, e.g. verbs with different valency patterns (transitive, intransitive); descriptor and classifier adjectives; adverbs of manner, degree, time; abstract and concrete nouns.

Different frequency bands were set considering the size of the corpus: low frequency lemmas were considered those with frequency between 500 and 1000 (between 9.34 and 18.69 occurrences per million words); mid frequency lemmas were those with frequency between 3000-5000 (between 56.10 and 93.40 occurrences per million words); and words with frequency of more than 5000 (93.40 occurrences per million words) were considered high frequency lemmas. Then, for nouns, adjectives and verbs, we selected 50 high frequency lemmas, 15 mid frequency lemmas, and 10 low frequency lemmas, i.e. 75 lemmas per word class. For adverbs, we selected 45 lemmas (30 high frequency, 10 mid frequency, and 5 low frequency lemmas).

For the evaluation of the sketch grammar, we used the following procedure:

1. Make a word sketch for one of the lemmas from the list;
2. Examine each gramrel in the word sketch, following these steps:

    a) When longest commonest match (LCM, Kilgarriff et al. 2015) is displayed – lines in grey under the collocates (see Figure 1 in Section 3.2); they are the most common realisation of the collocation in the corpus - , check if the collocation seems good[12] (in this way, we also checked the usefulness of information in LCM);

    b) Examine the list of collocates and determine if most of them seem good;

    c) examine the first 20 concordances of each of the top 25 collocates;[13]

    d) if bad matches are found, examine more concordances;

[12] In the paper "A Quantitative Evaluation of Word Sketches" (Kilgarriff et al. 2010), human experts (linguists and lexicographers) were asked to assess collocations and determine whether collocates were "Good; Good but wrong grammatical relation or POS-tagging error; Maybe (not striking collocate); Maybe (specialised vocabulary), and Bad" (ibid.: 376). We followed the same categorisation in our evaluation.

[13] Collocates were sorted by salience score, that is, by the strength of the collocation. Minimum collocate frequency was 3.

3. Consider the ratio between good and bad matches and:

    a) if there are many more good matches than bad ones, consider the gramrel good;

    b) if there are many bad matches, take note of the errors.

On the one hand, the evaluation corroborated the effectiveness of a handful of gramrels; on the other, it indicated that some of them performed less efficiently than expected, especially due to tagging errors.

As it was out of the scope of our research to work on the improvement of the tagger, we decided to try to find ways to work around some of those errors in order to improve the accuracy of some of the gramrels affected. We present two cases of adjustments of different nature: the first one related to the tokenisation of verbs with the particle *se*, and the second one related to the lack of tagging of participles as adjectives.

The particle *se* has many uses in Portuguese, thus its importance: 1. as a personal pronoun: reflexive pronoun, object; reflexive pronoun, indirect object; reflexive pronoun, object of reciprocal verbs; reflexive pronoun, object of infinitive; passive voice; unknown subject; expletive; part of verb expressing feelings, change of state, movement (Cegalla 2008: 562-563); 2. as a conjunction. It is known that those uses can be clearly determined by word sketches from dependency-parsed corpora. However, if this particle is correctly tagged, sketch grammar based on regex over POS-tagged corpora allows lexicographers to interpret its uses by analysing good concordances which reflect typical patterns.

Unfortunately, this was not the case with Freeling 3.0 for Portuguese. These are the problems involving *se* that have been found in CoPEP:

a) *Se* is tagged as a personal pronoun (PP3CN000) when it is actually a conjunction:

    Queremos saber se/se/PP3CN000 a inserção de ociosidade

nessa dada seqüência pode promover uma diminuição no valor da função-objetivo.

b) *Se* is tagged as a proper noun due to a capital letter;

c) Most of the time, *se* is not tokenised when postponed (thus connected to the verb by hyphen). Instead, it is considered a unit with the verb, forming the lemma verb+*se*.

Verbs matched are inflected for mode, tense, person, and number:

- verb modes: indicative, subjunctive, imperative, infinitive, and gerund
- verb tenses: present, imperfect, future, past, conditional, pluperfect
- person: 3rd person
- number: singular, plural
- gender: 0 (non-specified attribute; only for participle)

Examples:

Present: *deve-se* /VMIP3S0+PP3CN000/dever+se

Past: *desenvolveu-se* /VMIS3S0+PP3CN000/desenvolver+se

d) Less frequently, *se* is tokenised, lemmatised and tagged as a personal pronoun. In those cases, the hyphen is also tokenised and tagged as such (Fg). For example, the word form *escolhe-se*:

escolhe /VMIP3S0/escolher - /Fg/- *se* /PP3CN000/se

e) Since *se* is not tagged when it is part of verb+*se* lemma, it is ignored for the analysis of the following *se*, which is tagged as a pronoun and not as a conjunction:

Verifica-se se a empresa… ('it is verified if the company…')

verifica-se /VMIP3S0+PP3CN000/verificar+se *se* /PP3CN000/se

a /DAoFSo/o *empresa* /NCFSooo/empresa

The most significant problem to be tackled is the lack of capturing the use of *-se* when a verb is searched for in the word sketch function. This means that although a verb can occur with or without *-se*, there is no way to find such occurrences because the instances of the verb followed by *-se* are never matched.

Many different queries have been written to overcome this problem with the pronoun *se*, and all of them failed. After describing the problem to the Sketch Engine support team and showing the different workaround attempts, they proposed a reconfiguration of the Portuguese pipeline to accommodate our needs. A new corpus template - "Freeling Portuguese DEVELOPMENT" was created. Besides "lempos", "lc", and the three ordinary attributes [word, tag, lemma], three respective multi-value attributes [morphs, tags, morphemes] were added to the corpus. The attribute "morphemes" was created to account for verbs with clitics: it contains the lemma of the verb and all the pronouns (corresponding to what was joined by the "+" sign in the old pipeline). Morphological tags for the parts comprise "tags" and just the parts of the wordform separated by hyphens are "morphs", i.e. for verbs with clitics, this attribute can be the verb-stem part, the forms of the pronouns, and the suffix.

The second adjustment performed on the queries concerned the fact that the fix found for the lack of tagging participles as adjectives ended up causing a series of other problems. The original workaround consisted in adding the tag V.P.* for adjectives, as in this gramrel:

```
*DUAL

=Adj-Part mod %w_N/%w_Adj-Part mod N

2:"A.*|V.P.*" 1:"N.*"
```

As expected, the gramrel finds good collocations like *elevado teor* (lit. 'raised level'), where the participle form is an adjective that typically collocates with

the noun *teor*. Without the addition of V.P.*, collocations like this one would not have been found. Nevertheless, verbs *ser*[14] ('to be'), *ter*[15] ('to have') and *haver*[16] ('to have') are primary verbs, i.e. "can function as both auxiliary and main verbs" (Biber et. al 2015: 104). Thus, when they precede the structure V.P.*+N, in the vast majority of cases the participle form functions as a verb, not as an adjective. *Ser* makes up passive structures when followed by a participle verb form, while *ter* and *haver* followed by a participle verb form indicate a compound form with tense and aspectual functions.

For those situations, we had to come up with amendments to make sure that the gramrel matched only participle forms functioning as adjectives. Below, we touch on the problems that have arisen due to the addition of V.P.* to the query 2:"A.*|V.P.*"1:"N.*, and solutions proposed. Each of the three verbs was dealt with separately and solutions were put together in the end to form the final query.

The combination of verb *ser* + V.P.* + noun forms a passive structure, e.g. *são apresentados resultados* (lit. 'are presented results'). Thus, we first defined that *ser* cannot precede V.P.*: `[lemma!="ser"]"R.*"?2:"V.P.*"1:"N.*"`.[17] However, not all verb forms of the verb *ser* were captured by that query due to a lemmatising error in Freeling. As verbs *ser* and *ir* ('go') have the same forms in the simple past and in the third person plural of pluperfect, Freeeling 3.0 maps them all to verb *ir* only. Consequently, another workaround was needed to fix this problem: the creation of a rule with which any item can be met but those verb forms.

For the cases where *ser* is a copular verb, we performed a CQL search for this lemma (and the word forms mentioned above) followed by the participle forms

[14] *Ser* as a main verb is a copular verb, i.e., it is used "to associate an attribute with the subject of the clause" (Biber et al. 2015: 140).
[15] As a main verb, *ter* refers to the idea of possession, family connections, composition, etc.
[16] As a main verb, *haver* means 'there to be'.
[17] An optional adverb ("R.*"?) was included between each one of the auxiliary verbs above and the participle form in order to increase the accuracy of matches.

of six verbs (*elevar*, 'to raise'; *determinar*, 'to determine'; *limitar*, 'to limit'; *variar*, 'to vary'; *reconhecer*, ' to recognise'; *moderar*, 'to moderate') that have appeared as participial adjectives qualifying objects of the verb *ter*. There were only 47 occurrences in the whole 40-million-word corpus, and in only nine of them the verb *ser* was acting as a copular verb.

Despite the well-known existence of other verbs besides the ones investigated whose participle forms can also act as adjectives, the analysis of the sample verbs indicates that the occurrences of participial forms as prenominal adjectives in noun phrases in predicative function, whose linking verb is *ser*, have very low frequency when compared to the number of passive structures realised by the same word forms, i.e. *ser* acting as an auxiliary verb, participle form as a main verb, and a noun as agent of the passive.

For the compound structure *ter* + V.P.* + noun which indicates tense and aspect, such as in *tem apresentado resultados* ('has presented results'), the solution was to negate *ter* from the structure - `[lemma!="ter"] "R.*"?2:"V.P.*" 1:"N.*"`- in the same way as with the verb *ser* above. However, by not matching *ter* at all, occurrences of the verb acting as a main verb are not found, that is, collocations formed by participial adjectives + nouns (e.g. *elevada densidade*, lit. 'raised density') are not found when following *ter*. As sample analysis of the verb *ter* followed by *elevar* (V.P.*) (whose participial form is *elevado*)[18] showed a surprising 100% occurrence rate of *ter* acting as a main verb, which raised important questions: taking into account this apparent collocate loss, would that be a considerable problem? What about other participial adjectives that would not be captured; would we miss a great deal of relevant information?

Firstly, we conducted an analysis on the example of *elevar* (V.P.*) + noun which revealed only 1.14% of occurrences of such collocations with *ter*. Next, we

[18] In Portuguese, lemmas of (participial) adjectives are in singular, masculine form. When in use, adjectives take on number and gender inflections.

counted the total number of occurrences of *ter* + V.P.* + noun in CoPEP and manually analysed a random sample of 10% of concordances. Out of such sample, only 2.18% of occurrences corresponded to instances of *ter* as a main verb. Interestingly, *elevado* was the most frequent adjective found, followed by the other five adjectives with very low frequencies in the sample. These results indicated that negation of *ter* in the query would not significantly affect the analysis.

Nevertheless, further analyses were carried out in order to confirm such finding. This time, the other five participial adjectives identified in the previous analysis were looked at in more detail. We compared the number of times each of them collocated with nouns in structures preceded and not preceded by *ter* and found out that frequencies of *ter* + V.P.* + noun were always much lower than frequencies of their counterparts.

All the tests above led us to the conclusion that the option to miss some collocates for the benefit of better pattern matching seemed to be a reasonable trade-off.

Finally, we present the case of the compound *haver* + V.P.* + noun, which has the same function as the verb *ter*. Once again, the solution was to avoid matching *haver* by negating it in the query. In fact, *haver* is used much less frequently than *ter* in compound constructions, which means that any occasional loss in wrongly capturing participles used as adjectives instead of verbs would not be statistically relevant in the first place. Still, it is possible to reduce such an error by negating the verb *haver* in its plural form from the query. This is because the verb *haver* as a main verb means 'there to be' and is an impersonal verb, i.e. it only occurs in third person singular. That rule excludes almost half of the occurrences of *haver* in such structure. Despite the possibility of capturing haver as a main verb among the remaining concordances, the number of potentially missed combinations of participial adjectives and nouns is negligible in comparison with the total amount of such

combinations in the corpus, totalling only 0.7%.

All the solutions proposed above had to be merged in a single query to yield the correct results. These are the gramrels and queries for matching (participial) adjective + noun:

```
*DUAL

=Adj-Part mod %w_N/%w_Adj-Part mod N

2:"A.*"1:"N.*"

[lemma!="ser|ter" & !(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"][lemma!="ser|ter" &
!(lemma="haver"&tag="VM.*P.")&
lc!="foi|foram|fui|fomos|foste"]"R.*"?2:"V.P.*"1:"N.*"
```

It should be noted that there is a limit to how much effort can be put into finding workarounds for POS-tagging errors in our new sketch grammar definitions. Firstly, we must consider the fact that this sketch grammar is just one requirement for the development of a larger project, i.e. conceptualising and compiling a model for a dictionary of academic Portuguese. Secondly, making amendments is not the solution; to definitively overcome such limitations, the tagger should be improved. But that is one important lesson to be taken from this process, namely that the quality of information provided to lexicographers, in this case through word sketches, relies not only on definitions of grammatical relations in the sketch grammar, but also on the accuracy of tools such as taggers or parsers, and also on the quality of corpus data.

4 SUMMARY AND CONCLUSION

The sketch grammar for academic Portuguese, developed for exploration of grammar and lexis of Portuguese in the CoPEP corpus, has had implications not only for our work on the dictionary of Portuguese for university students, but also for Portuguese corpora in general. A comparison with the default sketch grammar available for Freeling-tagged corpora of Portuguese reveals

that AcadPortSkG comprises a larger number of grammatical relations for nouns, verbs and adjectives, and completely new rules for adverbs, thus broadening word class coverage. In addition, the queries of existing sketch grammars, which were used in developing AcadPortSkG, were adapted and now yield better results. Lastly, AcadPortSkG contains queries which were carefully devised in a way to overcome detected annotation errors, making it more accurate.

This new grammar, with broader coverage and more complex gramrels, yields very rich results - most of the times, it produces more data than can be handled by a human, which is in fact typical of sketch grammars for automatic extraction of lexical data (Kosem et al. 2013). We have already successfully conducted the first tests of automatic extraction, both on the entire corpus, as well as on both subcorpora of the two varieties of Portuguese. The proposed dictionary of academic Portuguese can now take maximum advantage of this procedure, meaning that lexicographers can get vast amounts of structured information in the dictionary-writing system. The richness of the lexicographic evidence obtained from the corpus due to the underlying sketch grammar is thus manageable, enabling compilation of more accurate dictionary entries.

What is more, AcadPortSkG can also be used with any other corpus of Portuguese tagged with Freeling 3.0. Such corpora will benefit from our sketch grammar on two levels. In terms of their exploration, the users of the corpora will be able to conduct a more thorough and reliable lexical analysis due to a greater number of grammatical relations and their (improved) accuracy. Concerning the development of tools for Portuguese, the very process of sketch grammar development has revealed problems with corpus annotation that can be used to improve the Freeling tagger and inform other resource developers of potentially problematic areas. One of such improvements has already been implemented by the Sketch Engine team. Our identification of the manner that Freeling 3.0 tokenised and lemmatised verbs with *-se*, together with a thorough description of annotation implemented and detailed explanation of that

particle function led to the reconfiguration of the Portuguese pipeline in the system.

There is still plenty of room for improvement, and we provide a few suggestions here: a) use of macros in the language m4. Macros are used to avoid repetition of recurring patterns. For instance, a macro "adjective" can be defined that includes adjectives and participle forms, thus "A.*|V.P*" does not have to be written every time adjectives are included in queries; b) improvement of corpus annotation, as exemplified in our case on *–se*; c) enrichment of the sketch grammar by devising queries for grammatical relations that are currently not covered.

Although this sketch grammar can be further improved, the current version already yields very good results, for both academic and general Portuguese. Thus, we decided to make it available in the Sketch Engine for researchers using/making Freeling-tagged corpora of Portuguese.

ACKNOWLEDGMENTS

This paper has been supported by the Coordination for the Improvement of Higher Education Personnel (Capes, *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) – Brazil with a Scholarship for the first author (n. 0973/13-0), and by the Slovenian Research Agency, infrastructure programme Centre for Applied Linguistics (I0-0051). The authors would also like to thank ENeL (ISCH COST Action IS1305) for the Short-term Scientific Mission grant.

REFERENCES

Araneum Portugallicum Maius. Available at: https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/pt_araneum_maius (Accessed on 23 November 2016).

Atkins,S. B.T., and Rundell, M. (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Benko, V. (2014a): Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.): *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655*: 257–264. Brno: Springer International Publishing Switzerland.

Benko, V. (2014b): Compatible Sketch Grammars for Comparable Corpora. In A. Abel, C. Vettori, and N. Ralli (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*: 417–430. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.

Benko, V.: *Araneum Portugallicum Maius, verze 15.05*. Ústav Českého národního korpusu FF UK, Praha 2015. Available at: https://kontext.korpus.cz/first_form?corpname=aranea%2Faranport_pt_ar13__b_a# (Accessed on 23 November 2016).

Biber, D., Conrad, S., and Leech, G. (2015) : *Longman Student Grammar of Spoken and Written English*. Harlow: Pearson Education.

Bick, E. (2000): The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Arhus: Aarhus University Press.

Capes. Available at: www.capes.gov.br (Accessed on 4 February 2016)

Capes' Areas of Knowledge Classification. Available at: http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/tabela-de-areas-do-

conhecimento-avaliacao (Accessed on 4 February 2016).

Cegalla, D. P. (2008): *Novíssima gramática da língua portuguesa*. São Paulo: Ed. Nacional.

CLUL (Centro de Linguística da Universidade de Lisboa). Online Resources. Available at: http://clul.ul.pt/en/resources (Accessed on 20 November 2016)

Corpus Brasileiro. Available at: http://corpusbrasileiro.pucsp.br/cb/Acesso.html (Accessed on 20 November 2016).

Corpus do Português: genre/historical. Available at: www.corpusdoportugues.org/hist-gen/ (Accessed on 20 November 2016).

CRPC - Corpus de Referência do Português Contemporâneo. Available at: http://alfclul.clul.ul.pt/CQPweb/crpcfg16/ (Accessed on 23 November 2016).

Gantar, P., Kosem, I., and Krek, S. (2016): Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29 (2): 200–225.

Généreux, Michel, Iris Hendrickx, and Amália Mendes (2012): Introducing the Reference Corpus of Contemporary Portuguese On-Line. *Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC 2012*: 2237-2244. Istambul.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013): The TenTen Corpus Family. *Proceedings of the 7th International Corpus Linguistics Conference*: 125–127. Lancaster.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., and Viks, Ü. (2015): Automatic generation of the Estonian Collocations Dictionary database. In I. Kosem, M Jakubíček, J. Kallas, and S. Krek (eds.): *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United*

Kingdom: 1-20. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing.

Kilgarriff, A., and Kosem, I. (2012): Corpus tools for lexicographers. In S. Granger, and M. Paquot (eds): *Electronic Lexicography*: 31–55. Oxford: Oxford University Press.

Kilgarriff, A., Baisa, V., Rychlý, P., and Jakubíček, M. (2015): Longest-commonest Match. In I. Kosem, M Jakubíček, J. Kallas, and S. Krek (eds.): *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom:* 397–404. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing.

Kilgarriff, A., Kovář, V., Krek, S., Srdanovic, I., and Tiberius, C. (2010): A Quantitative Evaluation of Word Sketches. In A. Dykstra, and T. Schoonheim (eds.): *Proceedings of the XIV Euralex International Congress*: 372–379. Leeuwarden: Fryske Akademy; Afûk.

Kilgarriff, A., Rychlý, P., Smrz, P., and Tugwell, D. (2004): The Sketch Engine. In G. Williams, and S. Vessier (eds.): *Proceedings of the 11th EURALEX International Congress*: 105–115. Lorient: Universite´ de Bretagne-Sud, Faculte´ des lettres et des sciences humaines.

Kosem, I., Gantar, P., and Krek, S. (2013): Automation of lexicographic work: an opportunity for both lexicographers and crowdsourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik (eds.): *Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*: 32–48. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Kosem, I., Gantar, P., Logar, N. and Krek, S. (2014): Automation of

lexicographic work using general and specialized corpora: two case studies. In A. Abel, C. Vettori, and N. Ralli (eds.): *Proceedings of the XVI EURALEX International Congress: The User in Focus*: 355–364. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism.

Kuhn, T.Z., and Ferreira, J.P. (2016): Building a corpus of written academic texts in Portuguese. *Teaching and Language Corpora Conference (TaLC12). Book of* Abstracts: 103. Giessen.

Linguateca. Available at: http://www.linguateca.pt/ (Accessed on 20 November 2016)

Logar, N., and Kosem, I. (2013): TERMIS: a corpus-driven approach to compiling an e-dictionary of terminology. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik (eds.): *Electronic Lexicography in the 21st Century: Thinking Outside the Paper: Proceedings of the eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*: 164–178. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Newspapers in Portuguese (CetemPúblico, CetenFolha). Available at: https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/portuguese (Accessed on 28 November 2016) .

NILC (Interinstitutional Center for Computational Linguistics). Tool and Resources. Available at: http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources (Accessed on 20 November 2016)

Oxford Portuguese Dictionary (2015). S. Lopez, A. Frankenberg-Garcia, and H. Newstead. Oxford: Oxford University Press.

Padró, L., and Stanilovsky, E. (2012): FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*: 1–7. Istanbul.

Peixoto, R. M. T. (2015): *O Fenômeno (De)Queísta no Corpus do Português Brasileiro Acadêmico*. Unpublished Master's Degree Dissertation. Porto Alegre: PUCRS.

Perini, M. A. (2002): *Modern Portuguese: A reference grammar*. New Haven: Yale University Press.

Portuguese Web 2011 (ptTenTen11, Palavras parsed). Available at: https://the.sketchengine.co.uk/bonito/corpus/first_form?corpname=preloaded/pttenten11 (Accessed on 6 April 2016).

Portuguese Web 2011 (ptTenTen11, Freeling v3). Available at: https://the.sketchengine.co.uk/bonito/corpus/corp_info?corpname=preloaded/pttenten11_freeling_v3_1 (Accessed on 23 November 2016).

Rundell, M., and Kilgarriff, A. (2011): Automating the creation of dictionaries: where will it all end?. In F. Meunier, S. De Cock, G. Gilquin, and M. Paquot (eds.): *A Taste for Corpora: In honour of Sylviane Granger*. Amsterdam: John Benjamins.

Scielo Brazil Analytics. Available at: http://analytics.scielo.org/w/publication/article?collection=scl (Accessed on 24 November 2016).

Scielo Brazil. Available at: www.scielo.br (Accessed on 15 February 2016)

Scielo Portugal Analytics. Available at: http://analytics.scielo.org/w/publication/article?collection=prt (Accessed on 24 November 2016).

Scielo Portugal. Available at: www.scielo.mec.pt (Accessed on 1 February 2016)

Scielo. Available at: www.scielo.org (Accessed on 23 November 2016)

Sketch Engine. Available at: https://www.sketchengine.co.uk (Accessed on 20 November 2016)

IZDELAVA SLOVNICE BESEDNIH SKIC ZA AKADEMSKO PORTUGALŠČINO

Prispevek predstavlja izdelavo nove slovnice besednih skic, ki je bila izdelana posebej za CoPEP, 40-milijonski korpus besedil iz znanstvenih revij. Korpus je bil označen z označevalnikom Freeling v3, privzetim označevalnikom v orodju Sketch Engine za korpuse portugalščine. Najprej na kratko predstavimo korpus CoPEP, razloge za njegovo izdelavo, podprte s pregledom obstoječih korpusov portugalščine. Sledi pregled in evalvacija obstoječih slovnic besednih skic za portugalščino, katere glavni zaključki so, da so obstoječe slovnice besednih skic zelo pomanjkljive in potrebne številnih popravkov in dopolnitev. Čeprav so nekatere poizvedbe slovničnih relacij ali njihovi deli koristni, pa je bilo, tudi z vidika naših raziskav, veliko smortneje izdelati povsem novo slovnico besednih skic. Osrednji del prispevka je tako posvečen podrobnemu opisu izdelave nove slovnice besednih skic in predstavitvi nekaterih težav, s katerimi smo se srečali. Največje težave pri pripravi poizvedb v slovnici besednih skic so povzročale določene pomanjkljivosti označevalnika, kar smo skušali rešiti z dodajanjem številnih pogojev v poizvedbe. Pri nekaterih težavah, kot je npr. označevanje osebnega zaimka *se,* smo rešitev iskali zunaj slovnice besednih skic in s pomočjo avtorjev orodja Sketch Engine izboljšali postopek priprave vseh korpusov portugalščine, ki so označeni z označevalnikom Freeling v3. Prispevek zaključimo s povzetkom glavnih ugotovitev, izpostavimo pomen rezultatov za nadaljnje raziskave, navedemo pa tudi nekaj predlogov za nadaljnjo izboljšavo slovnice besednih skic. Pomembna ugotovitev je, da besedne skice, izdelane na podlagi nove slovnice besednih skic, ponujajo natančnejše in precej bogatejše rezultate od tistih, ki jih dobimo z uporabo trenutno privzete slovnice besednih skic za portugalščino v Sketch Enginu. Zaradi bogatosti informacij lahko novo slovnico besednih skic uporabimo tudi za napredne leksikografske namene, kot je npr. avtomatsko luščenje leksikalnih podatkov iz korpusa CoPEP; to metodo bomo namreč uporabili pri izdelavi predlaganega slovarja portugalščine za študente. Čeprav je bila izdelana predvsem za namene akademske portugalščine, je slovnica besednih skic dragocen nov vir za leksikografske in korpusne raziskave portugalskega jezika, saj se jo lahko uporabi na vsakem korpusu, označenem z

označevalnikom Freeling v3. Posledično smo omogočili uporabo slovnice besednih skic vsem uporabnikom orodja Sketch Engine.

Ključne besede: slovnica besednih skic, portugalščina, korpus, slovar, evalvacija

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna.

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International.

https://creativecommons.org/licenses/by-sa/4.0/