LeMe–PT: A Medical Package Leaflet Corpus for Portuguese

Alberto Simões
2Ai, School of Technology, IPCA, Barcelos, Portugal

Pablo Gamallo
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
University of Santiago de Compostela, A Coruña, Spain

Abstract

The current trend on natural language processing is the use of machine learning. This is being done on every field, from summarization to machine translation. For these techniques to be applied, resources are needed, namely quality corpora. While there are large quantities of corpora for the Portuguese language, there is the lack of technical and focused corpora. Therefore, in this article we present a new corpus, built from drug package leaflets. We describe its structure and contents, and discuss possible exploration directions.

2012 ACM Subject Classification Computing methodologies → Information extraction; Computing methodologies → Language resources

Keywords and phrases drug corpora, information extractiom, word embeddings

Digital Object Identifier 10.4230/OASIcs.SLATE.2021.10

Supplementary Material *Dataset*: `https://github.com/ambs/LeMe`

Funding This project was partly funded by the project "NORTE-01-0145-FEDER-000045," supported by Northern Portugal Regional Operational Programme (Norte2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER), by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the project "UIDB/05549/2020", and through the IACOBUS program, managed by GNP and AECT. In addition, it has received financial support from DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), eRisk project (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08, Groups of Reference: ED431C 2020/21, and ERDF 2014-2020: Call ED431G 2019/04) and the European Regional Development Fund (ERDF).

1 Introduction

Drug Package Leaflets (DPL), also known as Patient Information Leaflets (PIL), are documents that contain valuable information for patients about the characteristics of medicines. Each DPL provides useful information about a drug, mainly stating the active substance that constitutes the drug, listing side effects, describing interactions with other drugs, and describing the drug's safety and efficacy, among other information.

In Portugal, DPLs are publicly accessible on the web, through the Portuguese National Authority of Medicines and Health Products website (Infarmed). This includes the documentation for all drugs currently accepted in the country, as well as some others that were previously approved but were late removed from the market.

Given the free nature of these documents and their relevant terminological content from different perspectives (drugs, illnesses, secondary effects), make this information highly valuable for Natural Language Processing (NLP), to be applied in different tasks, as information extraction, question answering solutions or machine translation, just to mention a few.

This contribution describes the creation of the LeMe–PT Corpus (Leaflets of Medicines), a corpus comprising more than a thousand of DPLs, and a set of experiments, including relation extraction and word similarity, performed to evaluate the relevancy of the corpus contents. In the next section related projects and works are introduced. Section 3 describes the corpus construction and its contents, and Section 4 presents some experiments for information extraction, and word embeddings creation and validation. Finally, we conclude with some future directions for the corpus use.

2 Similar Resources

In this section, we focus on describing some related work, including both other works where DPLs were used for corpus building, as well as other medical corpora, that were used for linguistic analysis and information extraction.

The EasyLecto system [17] aims to simplify DPLs by replacing the technical terms describing adverse drug reactions with synonyms that are easier to understand for the patients. EasyLecto simplifies the text of DPLs so as to improve their readability and understandability, and thereby allowing patients to use medicines correctly, increasing, therefore, their safety. This work is based in the Spanish version of the MedLinePlus Corpus.[1] The authors designed a web crawler to scrape and download all pages of the MedLinePlus web containing information on drugs and related diseases and, finally, stored the content into JavaScript Object Notation (JSON) documents. In order to evaluate the quality of their system, 306 DPLs were selected and manually annotated with all adverse drug reactions appearing in each document. The main problem of this approach is that it relies on external terminologies to provide synonyms. To overcome this limitation, the same authors described a new method in a more recent work [16], based on word embedding, to identify the more colloquial synonym for a technical term.

A similar work on readability of DPLs was previously reported [10]. In this work, the authors built a medical thesaurus of technical terms appearing in these documents, aligning them with a colloquial equivalent, easier to understand for the patients, in order to substitute the technical names by their colloquial equivalents, and making the DPLs easier to read and understand.

There are also corpus-based studies focused on the analysis of linguistic phenomena in DPLs. In [22], the author aimed at identifying keywords and frequent word sequences (recurrent n-grams) that are typical of this type of text. It was used a corpus with 463 DPLs and 146 summaries of drug characteristics written in English. Similar corpus-driven studies for Polish [23] and Russian [24] were reported. In [14], the building of a corpus comprised of medicine package leaflets medicines is described as well as its use as a teaching resource for Spanish–French translation in the medical domain.

There is very little literature on information extraction from DPLs, [1] is a Master's thesis work that describes a system to automatically extract entities and their relationships from Portuguese leaflets, with the aim to get information on dosage, adverse reactions, and so on. This work applies Named Entity Recognition (NER) and Relation Extraction (RE) techniques on text, to populate a medical ontology. Unfortunately, neither the corpus, software or the extraction results are freely available.

[1] See `https://www.nlm.nih.gov/medlineplus/spanish/`.

We should also highlight works on information extraction from medical corpora, not necessarily from DPLs. In [15] the authors describe a method to extract adverse drug reactions and drug indications in Spanish from social media (online Spanish health-forum), in order to build a database with this kind of information. The authors claim that health-related social media might be an interesting source of data to find new adverse drug reactions.

Still in Spanish, [21] aims to develop tools and resources for the analysis of medical reports and the extraction of information (entities and terms) from clinical documents. Regarding semantic information, the work reported in [19] describes the steps to create in-domain medical word embeddings for the Spanish using FastText and both the ScioELO database and Wikipedia Health as source of information [12].

For Portuguese, there is work on extracting information from medical reports [3]. The main issue with this kind of system is the availability of such corpora, as the data protections laws require the anonymization of the documents and, even after that process, hospitals would not allow the public release of such a corpus.

Finally, we should point out that there are very few corpora based on compiling leaflets which are available for free exploitation. One of the few examples is the Patient Information Leaflet (PIL) corpus, which consists of 471 English documents in Microsoft Word and HTML formats[2].

In Portuguese, the *Prontuário Terapêutico Online* is a website that allows several types of search on the Infarmed database of drug leaflets[3]. This site also includes some extra information that is not present directly in the leaflets, but was added to help doctors on their drug prescription.

It is also relevant to mention the availability of generic medical corpora. One of the most known is the European Medicines Agency (EMA) Parallel Corpora, available from the OPUS project [20]. From this corpus a set of related projects were developed, as a Romanian corpus [8], or the organization of information extraction tasks under the Cross Language Evaluation Forum (CLEF), as described in [6] and [9].

3 The Corpus

Our corpus was built with drug package leaflets obtained from the Infarmed[4] website[5] Given the interactive process required to download these documents, the DPLs download was performed manually, using as seed a list of drug active substances. Each active substance was searched in the website, and a random drug package was chosen (it was not given any priority to generic or original drugs). When different pharmaceutical forms were available, the less common was chosen.

On some situations, the document linked from the Infarmed website is available at the European Medicines Agency website. In these situations, the documents include a full report on the drug, performed tests, effects, and so on. At the end, in an appendix, these reports include a copy of the DPL[6]. So, in these situations, the document was truncated to include only this specific appendix.

[2] See `http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL/`.
[3] Available at `https://app10.infarmed.pt/prontuario/`.
[4] Infarmed, available at `http://infarmed.pt`, is the Portuguese National Authority of Medicines and Health Products.
[5] Note that these documents are copyrighted by the respective pharmaceutical company. We are just easing the access to these documents in a textual format.
[6] Some of these reports include different DPLs copies, accordingly with the various drug dosages available in the market.

The corpus include 1191 different package leaflets, referring to 1191 different active substances (some leaflets refer to compound active substances). The majority of these documents are divided into five to seven different sections. The most common are:

- what is the drug application, including sometimes its type;
- precautions the patient should take before using the drug;
- the usual dosage, depending the illness, age and other patient characteristics;
- the possible secondary effects and/or interactions with other drugs;
- how to store the package and other less relevant information.

For the documents obtained from the European Medicines Agency, they were automatically cleaned, removing the introductory report. Some still include different variants of the instructions, that will require manual cleanup. At the current version (v1.0), the corpus comprises about 3 000 000 tokens, from which about 2 650 000 are words, accounting to over 30 000 different word forms.

The corpus is available in a text file for each specific active substance and it is minimally annotated with XML-like tags:

- Title tags, dividing the different sections of the document. Most documents include only the five or six sections. A few do not follow this specific structure, and have more than ten sections.
- Item and Sub-item tags, annotating all lists automatically detected in the document.

We intend that new versions of the corpus include further annotations, namely on illnesses, drugs, secondary effects, and other relevant information. The next section describes the first steps towards the inclusion of this kind of information in the corpus.

4 Experiments

In this section we present some experiments performed with this corpus, presenting some directions for information extraction.

4.1 Regular-Expression based Information Extraction

One first experiment was performed to extract information about what is each substance. For that, the first section of each document was processed, trying to find two different kinds of relations:

- Hyponymy: referring to the medicine type. Examples of detected types are presented on Table 1. This relation was obtained for 1058 different substances. The extraction of this information is performed by the use of the following regular expressions:

```
que é \s+ uma? \s+ ([^.]+)
que é pertence.*? \s+ (?:por|d[eao]s?) \s+ ([^.]+)
```

  Note that these two regular expressions are applied in context, meaning they will be only activated in the proper section of the document.
- Condition or illness the medicine is adequate for, as shown in Table 2. This relation was obtained for 979 different substances. Follows a pair of examples of the different regular expressions used to extract this information:

```
tratamento \s+ ((?:\S+ \s+)?) d[aoe]s? \s+ ([^.]+)
(?:indicado|usado|utilizado) \s+ (?:para|n[ao]s?) \s+ ([^.]+)
```

These relations can be extracted with reasonable recall and high precision as the vocabulary used in this kind of document is quite controlled and the syntactic structures are recurrent. This can be comparable to the language used by lexicographers [18]. For instance, the relations mentioned above are extracted using six simple regular expressions. However, in some cases, this technique is extracting large sentences which should be reduced and simplified. Combining these simple text-mining techniques with some basic natural language processing techniques would allow for more compact extractions and higher quality data.

Given the amount of different possibilities to make the results better, at the current stage it was not performed any evaluation on precision or recall for these methods. Nevertheless, the manual annotation for some of these properties is planned, so that the corpus can also be used as an information retrieval test set.

Table 1 Examples of hyponymy relations extracted from LeMe–PT.

zolmitriptano *medicamentos denominados de triptanos*
zolpidem *medicamento de administração oral* | *medicamentos ansiolíticos, sedativos e hipnóticos*
valproato semisódico *anticonvulsivante*
valaciclovir *medicamentos designados de antivirais*
toxina botulínica A *relaxante muscular utilizado para tratar várias condições no corpo humano*
tramadol + dexcetoprofeno *analgésico da classe dos anti-inflamatórios não esteróides*

Table 2 Examples of conditions or illnesses extracted from LeMe–PT.

zolmitriptano *depressão* | *tratar as dores na enxaqueca*
zotepina *esquizofrenia, que tem sintomas como ver, ouvir ou sentir coisas que não existem*
tansulosina *sintomas do trato urinário inferior causados por um aumento da próstata*
tapentadol *dor crónica intensa em adultos*
tribenosido + lidocaina *hemorroidas externas e internas*

4.2 Words proximity using Word Embeddings

Some experiments were performed using Word2Vec [11]. More precisely, the corpus was pre-processed with the `word2phrases` script [13], which is shipped with the `word2vec` code, to create multi-word expressions and extracted word embeddings with the `word2vec` program, by training both a continuous bag of word model (CBOW) and a Skip-gram model, for a window size of 10 words, 15 iterations, and 300 dimension vectors.

Table 3 presents proximity terms for a set of words. For the first example, *alprazolam*, the list includes mostly other soothing drugs. For the second, *palpitações* [palpitations] the results are different kinds of heart rates dysfunctions. Finally, in the third column, *sonolência* [somnolence], the proximity terms are related to mental status, from soothing to tremors.

Table 3 Proximity terms obtained by Word2Vec (CBOW model).

alprazolam		palpitações		sonolência	
triazolam	0.843 623	aceleração	0.872 615	tonturas	0.803 442
diazepam	0.775 389	batimento cardíaco	0.857 808	sedação	0.801 227
alfentanilo	0.768 538	palpitações cardíacas	0.856 427	letargia	0.784 619
temazepam	0.762 558	batimento cardíaco acelerado	0.856 367	confusão mental	0.781 599
amissulprida	0.753 863	taquicardia	0.855 876	vertigens	0.761 623
midazolam	0.742 607	ritmo cardíaco lento	0.851 237	ataxia	0.747 047
tranquilizante	0.733 400	frequência cardíaca lenta	0.847 746	tremores	0.742 320
clonazepam	0.716 460	batimento cardíaco rápido	0.845 664	tremor	0.742 008
sedativo	0.716 204	acelerado	0.842 884	nervosismo	0.739 591
brotizolam	0.716 096	ritmo cardíaco rápido	0.838 991	coordenação	0.737 030

4.3 Word Embeddings Evaluation

In order to perform a basic evaluation task on the quality of the word embeddings generated from LeMe–PT corpus, an intrinsic evaluation was built by making use of a specific word similarity task, namely the outlier detection task [2, 5]. The objective is to test the capability of the embeddings to generate homogeneous semantic clusters. It consists of identifying the word that does not belong to a semantic class. For instance, given the set of words

$$S = \{lemon, orange, pear, apple, bike\},$$

the goal is to identify the word *bike* as an outlier of the class of fruits. One of the advantages of this task is that it has high inter-annotator agreement as it is easy to identify outliers when semantic classes are clearly defined.

To evaluate our embeddings model with the outlier detection task, five medical classes were built. Each one consists of eight words belonging to a specific class and eight outliers which do not belong to that class. The five classes are *analgesics*, *antidepressants*, *autoimmune diseases*, *respiratory diseases* and *pharmaceuticals*. The first four were elaborated by consulting specialized medical websites and the last one by choosing the first 8 drugs (in alphabetical order) from Infarmed. To give an example, Table 4 depicts the elaborated class of *antidepressants*.

As in this example, the five classes and their corresponding sets of outliers are unambiguous, thus there is no fuzzy boundary between class elements and outliers.

The outlier metric is based on a specific clustering method, called *compactness score*. Given a set of word elements $C = \{e_1, e_2, \ldots, e_n, e_{n+1}\}$, where $e_1, e_2, \ldots, e_n$ belongs to the same semantic class and e_{n+1} is the outlier, the compactness score compact(e) of an element $e \in C$ is defined as follows:

$$\text{compact}(e) = \frac{1}{n} \sum_{\substack{e_i \in C \\ e \neq e_i}} \text{sim}(e, e_i) \tag{1}$$

Table 4 Class of antidepressants and set of outliers.

Antidepressants	Outliers
imipramina	abiraterona
clomipramina	serotonina
amitriptilina	insónia
desipramina	tremores
nortriptilina	paracetamol
fluoxetina	doença
paroxetina	farmácia
citalopram	carro

An outlier e is detected if the value of compact(e) is lower than the compact$(e_i), \forall e_i \in C$, the scores of the words belonging to the class C. Two specific evaluation metrics are used: *accuracy* measures how many outliers were correctly detected by the system divided by the number of total tests. In [2], the authors also define *Outlier Position Percentage* (OPP) which takes into account the position of the outlier in the list of $n + 1$ elements ranked by the compactness score, which ranges from 0 to n (position 0 indicates the lowest overall score among all elements in C, and position n indicates the highest overall score).

To compare the embeddings generated from LeMe–PT corpus with other models generated from generic Portuguese corpora, we also applied the outlier task to three pre-trained Embeddings from Inter-institutional Center for Computational Linguistics (NILC) [7]: NILC-Word2Vec, NILC-FastText, and NILC-Glove.[7]

The three models contain 300 dimensions. They were generated from a vast corpus with 1 395 926 282 word tokens, i.e. about 600 times larger than LeMe–PT. NILC-Word2Vec and NILC-FastText were trained with both the Skip-Gram and and Continuous Bag-Of-Words (CBOW) algorithms.

Table 5 shows the *accuracy* and *OPP* scores obtained with the compared embedding models with both Skip-Gram and CBOW algorithms (except for Glove as this is not applicable). The CBOW model derived from LeMe–PT clearly outperforms the other generic CBOW models. Concerning Skip-Gram, the differences are not so clear, although our model achieves the highest OPP value. These results suggest that the corpus built from DPLs is coherent and capable of generating competitive semantic models that are well adapted to the specific domain of medical leaflets.

5 Conclusions

In this document we present LeMe–PT, a corpus on drug package leaflets. It is freely available for download through a GitHub repository[8], and includes a larger number of documents in comparison with similar projects.

The documents are minimally annotated with a clear structure, allowing the extraction of information from different sections. Given its compilation process was manual and comprehensive, it includes drug leaflets from all medicine areas, and for every drug active substance currently in the Portuguese market[9].

[7] Available at `http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc`.
[8] Please visit `https://github.com/ambs/LeMe`.
[9] About twenty active substances included only one specific package, which did not have its leaflet available.

Table 5 Outlier Position Percentage (OPP) and Accuracy of several embedding models on the outlier detection dataset with 5 classes.

Model	OPP		Accuracy	
	skip-gram	cbow	skip-gram	cbow
LeMe–PT (word2vec)	**87.81**	**96.88**	62.50	**87.5**
NILC-word2vec	83.44	80.00	62.50	50.00
NILC-fasttext	83.44	78.75	62.50	42.50
NILC-glove	83.12		57.5	

While some first experiments on the relevance of the corpus were performed, there are different directions one can explore in the future:

- Further annotation of the documents, namely on illnesses and secondary effects. This process can be bootstrapped automatically, but a throughout manual validation would be imperative. This would allow the use of the corpus for the evaluation of information extraction tools.
- Combine the simple text-mining techniques with basic natural language processing techniques, in order to obtain higher quality data.
- Apply Named Entity Recognition (NER) techniques to extract relevant entities, and use these entities with Open Information Extraction techniques [4] to improve recall and extract more types of relationships between the different kind of entities present in these documents.
- The availability of these documents should be cross-country, meaning the possibility to create parallel or, at least, comparable corpora with a high degree of terminological information.
- Given the relevance of the SNOMED Clinical Terms[10] ontology, it would be interesting to perform a concept alignment with this structure, which could result on a bootstrap mechanism for a Portuguese subset of SNOMED CT.

References

1 Bruno Lage Aguiar. *Information extraction from medication leaflets*. PhD thesis, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal, 2010.

2 José Camacho-Collados and Roberto Navigli. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 43–50, Berlin, Germany, 2016.

3 Liliana Ferreira, António Teixeira, and João Paulo Silva Cunha. Medical information extraction in european portuguese. In *Handbook of Research on ICTs for Human-Centered Healthcare and Social Care Services*, pages 607–626. IGI Global, 2013. doi:10.4018/978-1-4666-3986-7.ch032.

4 Pablo Gamallo. An Overview of Open Information Extraction (Invited talk). In Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões, editors, *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 13–16, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/OASIcs.SLATE.2014.13.

[10] See https://snomed.org.

5 Pablo Gamallo. Evaluation of Distributional Models with the Outlier Detection Task. In Pedro Rangel Henriques, José Paulo Leal, António Menezes Leitão, and Xavier Gómez Guinovart, editors, *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*, volume 62 of *OpenAccess Series in Informatics (OASIcs)*, pages 13:1–13:8, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/OASIcs.SLATE.2018.13.

6 Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, João Palotti, and Guido Zuccon. Overview of the clef ehealth evaluation lab 2015. In Josanne Mothe, Jacques Savoy, Jaap Kamps, Karen Pinel-Sauvagnat, Gareth Jones, Eric San Juan, Linda Capellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 429–443, Cham, 2015. Springer International Publishing.

7 Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks, 2017. arXiv:1708.06025.

8 Radu Ion, Elena Irimia, Dan Ştefănescu, and Dan Tufiș. ROMBAC: The Romanian balanced annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 339–344, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/218_Paper.pdf.

9 Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João Palotti, and Guido Zuccon. Overview of the CLEF eHealth evaluation lab 2016. In *Lecture Notes in Computer Science*, pages 255–266. Springer International Publishing, 2016. doi:10.1007/978-3-319-44564-9_24.

10 Fabian Merges and Madjid Fathi. Restructuring medical package leaflets to improve knowledge transfer. In *IKE: proceedings of the 2011 international conference on information & knowledge engineering*, Las Vegas, Nevada, 2011.

11 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. cite arxiv:1301.3781. URL: http://arxiv.org/abs/1301.3781.

12 Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L18-1008.

13 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

14 Manuel Cristóbal Rodríguez Martínez and Emilio Ortega Arjonilla. El corpus de prospectos farmacéuticos como recurso didáctico en el aula de traducción especializada francés-español. In Vicent Montalt, Karen Zethsen, and Wioleta Karwacka, editors, *Current challenges and emerging trends in medical translation. MonTI 10*, pages 117–140. Universidad de Alicante, 2018.

15 Isabel Segura-Bedmar, Santiago de la Peña González, and Paloma Martínez. Extracting drug indications and adverse drug reactions from Spanish health social media. In *Proceedings of BioNLP 2014*, pages 98–106, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi:10.3115/v1/W14-3415.

16 Isabel Segura-Bedmar and Paloma Martínez. Simplifying drug package leaflets written in spanish by using word embedding. *Biomedical Semantics*, 8 (45), 2017. doi:10.1186/s13326-017-0156-7.

17 Isabel Segura-Bedmar, Luis Núñez-Gómez, Paloma Martínez, and M. Quiroz. Simplifying drug package leaflets. In *SMBM*, 2016.

18 Alberto Simões, Álvaro Iriarte, and José João Almeida. Dicionário-aberto – a source of resources for the portuguese language processing. *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence*, 7243:121–127, 2012. doi:10.1007/978-3-642-28885-2_14.

19 Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-1916.

20 Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

21 Pilar López Úbeda. Reconocimiento de entidades en informes médicos en español. In *Proceedings of Doctoral Symposium of the 33rd Conference of the Spanish Society for Natural*, 2018.

22 Łukasz Grabowski. Register variation across english pharmaceutical texts: A corpus-driven study of keywords, lexical bundles and phrase frames in patient information leaflets and summaries of product characteristics. *Procedia - Social and Behavioral Sciences*, 95:391–401, 2013. doi:10.1016/j.sbspro.2013.10.661.

23 Łukasz Grabowski. On lexical bundles in polish patient information leaflets: A corpus-driven study. *Studies in Polish Linguistics*, 9(1), 2014.

24 Łukasz Grabowski. Distinctive lexical patterns in russian patient information leaflets: a corpus-driven study. *Russian Journal of Linguistics*, 23(3):659–680, 2019. doi:10.22363/2312-9182-2019-23-3-659-680.