ISSN 1870-4069

A Proposal of Lexical Resources' Development for Ontological Learning in the Domain of Speech Disorders

Stephanie Vázquez, María Somodevilla, Ivo Pineda, Concepción Pérez de Celis

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla, Mexico

{stephanie.vazquez, mariajsomodevilla}@gmail.com,
{ivopinedatorres, mpcelish}@gmail.com

Abstract. Speech disorders in children are a condition that could reduce the opportunity to access education, health care and in the future could mean a worse socioeconomic outcome. Therefore, early diagnosis and timely therapy is really important to reduce their impact in later stages of life. This paper presents a method for the gathering of data for a corpus related to Speech Disorders in children; such corpus will serve as the base to generate a semi-automatic ontology intended as a tool for therapists to help in the diagnosis and shape up of a therapy strategy.

Keywords. Speech disorders, corpus building, crawling, dictionary building, semi-automatic ontology creation.

1 Introduction

A speech disorder is the difficulty to produce or to create the specific speech sounds to communicate. These disorders can range from simple sound substitutions to disability for understanding or using the language (motor-oral mechanism) for the speech. Causes could be as diverse as hearing loss, neurological disorders, brain injury, intellectual disability, or physical impairments as cleft lip [1].

According to Global Disability Rights 7.5% of the population in Mexico has some disability (about 9.17 million people) and 4.87% of people with disability has some type of speech disorder (0.45 million people). In kids and young people the speech disabilities are in some cases twice or four times higher than in adults [2]. The majority of people with disabilities do not have equal access to health care, education, and employment opportunities, do not receive the disability-related services that they require, and experience exclusion from everyday life activities, furthermore a disability is a development issue: evidence shows that persons with disabilities experience worse socioeconomic outcomes and poverty than persons without disabilities [3].

The importance of the early detection and diagnosis of a speech disorder abides in the social, economic and educative impact that such disorders have in the life of infants. Technology is used in order to assist in the process of diagnosis and treatment

of some speech disorders in children. ICT (Information and Communication Technologies) are helpful in almost every step to identify and provide a treatment for a speech disorder.

To deal with the problem of manipulating and organizing a big amount of data such as speech disorders' information the use of ontologies can be resorted to. An ontology provides through Semantic Web -an evolving extension of the World Wide Web- the semantics of information and services so that the Web can understand and satisfy requests for content made by people and machines [4]. Ontologies give an unambiguous and well defined structure for a clear and accurate representation of the data concerning a particular domain, in this case speech disorders, and thus, becoming a tool for diagnosis. Ontologies are made up of two main components: classes and relationships (See Fig. 1).

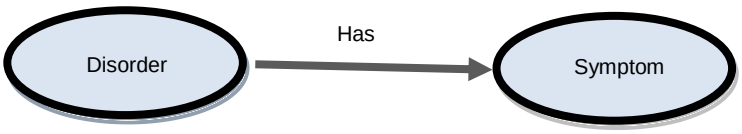

Fig. 1. Simple representation of the two main components in an ontology: classes and relationships.

Is proposed an ontology to organize and to look up the information relative to speech disorders such as different disorders, characteristics of each disorder, therapy plans, taxonomy of the speech disorders, and some other helpful information for the therapist and patient, as well as the relationships between all of them. One of the earlier steps in the development of an ontology is the conformation of a Corpus, in this case of documents relatives to the domain of speech disorders.

A Corpus is a large collection of texts. It is a body of written or spoken material upon which a linguistic analysis is based. The corpus may be composed of written language, spoken language or both. Spoken corpus is usually in the form of audio recordings. A corpus may be open or closed. An open corpus is one which does not claim to contain all data from a specific area while a closed corpus does claim to contain all or nearly all data from a particular field. Computer-processable corpora allow linguists to adopt the principle of total accountability, retrieving all the occurrences of a particular word or structure for inspection or randomly selected samples. Corpus analysis provide lexical information, morphosyntactic information, semantic information and pragmatic information [5].

This document is organized as follows: section 2 presents the state of the art through the discussion of some works related to the subject of the present work. Section 3 exposes the model proposed by the authors to build a corpus as a data source for the future ontology; subsections detail the construction of a very important component in this model: the *dictionary*. Section 4 presents the data obtained when testing the corpus with several algorithms and the resulting extended dictionary. Finally, in Section 5 the conclusions of applying the proposed model are outlined followed by the references.

2 State of the Art

Within the field of speech and language several works that use Information and Communication Technologies (ICT) have been conducted, focusing on some ailments such as dysphagia [6], on the automatic classification of the quality of pronunciation when treating disorders such as dyslalia or dysarthria [7], or an expert system for the initial evaluation of children with possible speech disorders [8]. A so-called ecosystem of smart ICTs that include electronic medical record management, standardized vocabularies, a knowledge database, ontologies for concepts within the domain of speech and language, and expert systems focused on supporting speech and language pathologists, doctors, students, patients, and their relatives can also be found [9]. There are also tools for the formation of professionals in the field of speech disorders based on ontologies and e-learning, which support future language therapists in their training process, as well as in their development of practical abilities [10]. Regarding language therapies, a mobile app that integrates therapy activities for children and that uses colloquial language, as well as games from the state of Chiapas, has been developed [11]. There even is a robust ontology that covers several aspects of speech and language therapies, with key concepts such as initial evaluation and patient profile, conducted tests, doctors and therapists catalog, list of disorders, speech and language fields, therapy and tracking plans and exercises, among others, that uses OpenEHR ontologies and constructs [12].

Regarding the semi-automatic creation of taxonomies for a given domain, several methods have been proposed that use techniques as diverse as formal Horn concepts and clauses analysis through logical inference validation [13], hierarchical clustering of documents based on sets of frequent concepts validated through prototype implementation [14], or a generalized algorithm of association rules that detects relationships between concepts and that detects the proper abstraction level for relationships definition [15].

Relevant to the building of corpus the main techniques have not varied a lot, and texts in a corpus need to be in electronic form. Thus, the fastest way to build a corpus is gathering data that is already digitalized or relying mainly in transcript into electronic form the audios, or documents [16].

In the present work, a method to gather information for the corpus building is proposed. This method also has the flexibility to feedback itself; once the initial dictionary is defined this can be updated with the extended dictionary obtained after completing the several steps into the method.

3 Information Retrieval Model for the Definition of Lexical Resources

In order to build the Corpus, it is necessary to gather a big amount of documents relevant to speech disorders through a Web Crawler. This crawler uses a predefined dictionary with some of the terms relevant to the domain. Once a representative amount

of those documents is obtained, they need to be pre-processed in several steps to clean up and standardize the data thought algorithms like normalization and stemming. Once the data is clean and ready to retrieve information some algorithms like word ranking and n-grams are applied to extend the original dictionary relevant to the domain of speech disorders. In this section each step in the conformation of a corpus will be explained.

The several steps in the task of building corpus and processing it can be seen in a diagram in Fig. 2.

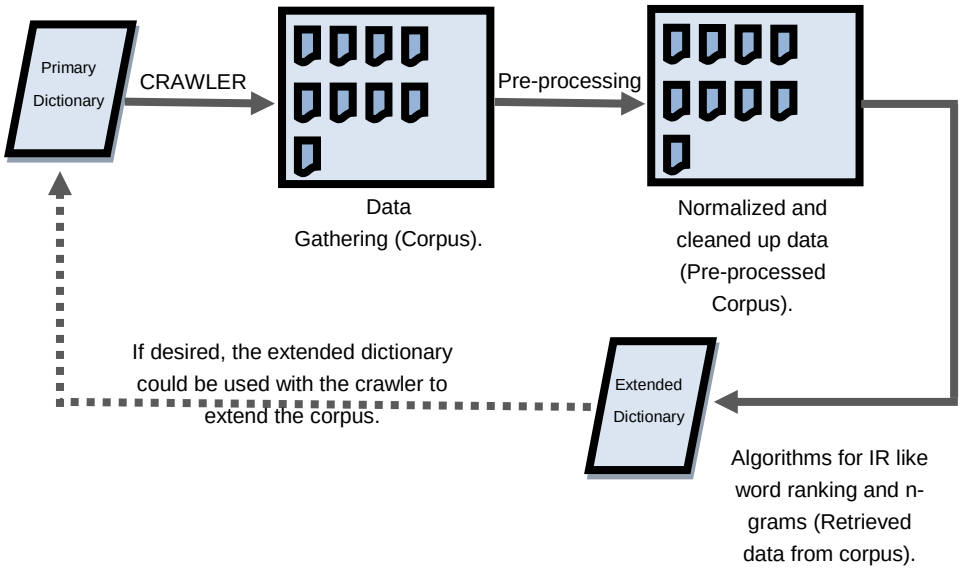

Fig. 2. Diagram of the steps to build and process a corpus.

3.1 **Corpus Creation**

The building of a corpus is divided into two stages: design and implementation. A good practice in the stage of design is to define what would ideally the corpus will have, in terms of the amount and the type of language, and then the parameters could be adjusted as the building goes along, keeping a careful record of what is in the corpus, so it can be added and amended later, and if others use the corpus they know what is in it [16].

In order to build a corpus there are a number of factors which need to be taken into consideration. These include size, balance and representativeness. The size of the corpus depends very much on the type of questions that are going to be asked of it.

The sample documents in our corpus would need to be balanced. Getting this balance right is not an exact science and there are no reliable ways of determining whether a corpus is truly balanced. One approach to achieving balance is to use an existing corpus as a model; research has suggested that samples of 2,000 to 5,000 words are sufficient.

A corpus can be said to be representative if the findings from that corpus are generalizable to language or a particular aspect of language as a whole. The notion of 'saturation' can be used. Saturation (at the lexical level) can be tested for by taking a corpus and dividing it into equal sections in terms of number of words. If another section of the same size is now added, the number of new items in the new section should be approximately the same as in the other sections [17].

The main tool to gather the information to build a corpus is a Web crawler. A crawler can be defined as an Internet *bot* that browses the World Wide Web, typically with the purpose of Web indexing. This crawler is fed with some initial *seed* pages to start its task. At their core is an element of recursion. They must retrieve page contents from an URL, examine that page for another URL, and retrieve that page, ad infinitum [18]. To find documents relevant to the domain, and not just a list of links and random data contained into the seed page, it is necessary to establish a primary dictionary at the beginning of the crawling.

3.2 **Dictionary Creation**

This dictionary is made of some of the more significative words into the domain. A simple way to identify these words is to take the domain taxonomy as a base to gather such list of words. The figure3 shows the taxonomy of speech disorders proposed by the DSM-5 manual of Mental Disorders [19].

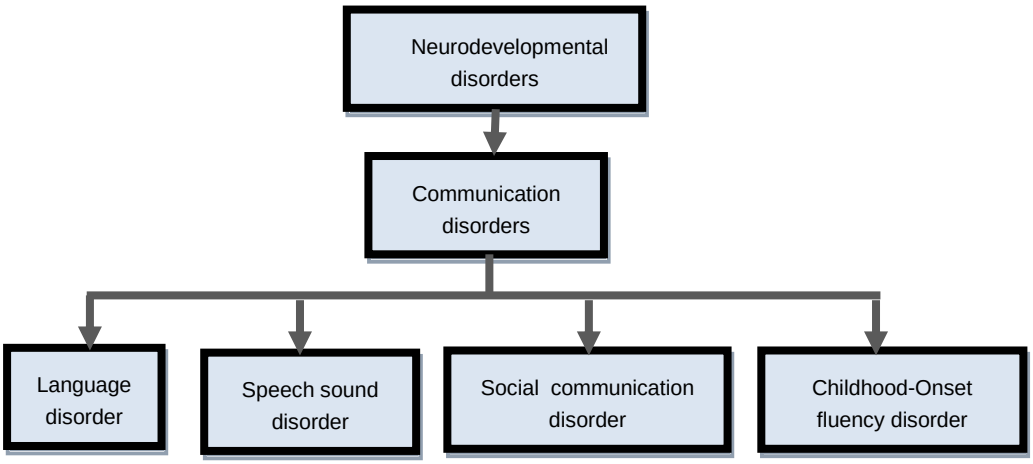

Fig. 3. Hierarchical Taxonomy of Speech Disorders according to DSM-5 manual.

Then the building of the primary dictionary to focus the results of the crawler can be started. As this taxonomy is a small one, the size of the dictionary using some other terms related to the ones included in our taxonomy could be increased. There are some other classifications for speech disorders that include specific names for each kind of *speech sound disorder*. In Fig. 4 another example of speech disorders classification can be seen.

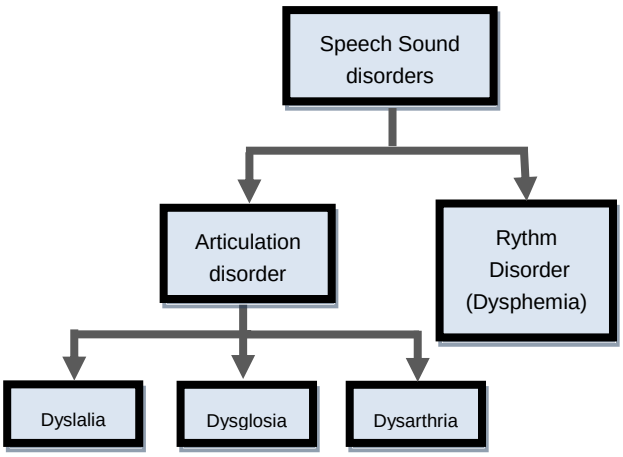

Fig. 4. Additional classification for speech sound disorders.

Since the ontology is focused only in the diagnosis and therapy of *speech sound* and *fluency* disorders are used just the relevant terms relative to such disorders. The table 1 shows the very first version of the primary dictionary.

Table 1. List of terms from the primary dictionary.

No.	Term(s)		Term(s)		Term(s)
1	Speech	6	Dysphemia	11	Rhythm disorder
2	Disorder	7	Speech sound disorder	12	Therapy
3	Dyslalia	8	Childhood-onset fluency disorder	13	Speech therapy
4	Dysglosia	9	Communication disorder	14	Logopedic therapy
5	Dysarthria	10	Articulation disorder	15	Speech development

Starting with the terms directly obtained from the branches of the taxonomies that are related with the domain, terms as *communication disorders, speech sound disorders, childhood-onset fluency disorder, articulation disorders, rhythm disorder, dyslalia, dysglosia, dysarthria, dysphemia* and every single significative word in those terms goes to the dictionary. As the ontology will also contain data about the therapies applied to the previously listed disorders it's also desirable to include related terms like *therapy, speech therapy, logopedic therapy* (the study and treatment of speech defects) and *speech development.*

The next step is the use of this dictionary to gather the corpus for the ontology using a web crawler written in Python language with the help of libraries *HTMLparser, urlopen* and *BeautifulSoup* [20]. Using some Web pages relevant to the domain of speech disorders (like *www.asha.org, medlineplus.com*, etc.) the traversing of those sites is started in search of each term of the dictionary once at a time and retrieving the data in each site visited, store that data in a file and then storing the links to and

visiting internal pages into the *seed page* provided like parameter to the crawler. An additional parameter for the crawler could be a maximum number of pages to visit in search of the term. After retrieving relevant data for all the primary dictionary terms the first version of our corpus is finished, but the processing of the corpus is not done.

3.3 **Data Preprocessing**

Pre-processing the data is the next step. This is done through several algorithms that normalize the texts contained in the corpus. Algorithms for removing escape characters, Unicode characters, punctuation marks, stop words, converting to plain text and capitalization are very useful to clean-up the data before being analyzed [21][22]. Again, with some Python routines are performed the algorithms to clean the data. Once all the data gathered into the corpus is normalized the next step in the process can be done.

In this step, information retrieval algorithms are implemented. Algorithms like word frequency and stemming are used [22]. After this last step a new list of terms for the extended dictionary is obtained. The more frequent terms found into the corpus are taken and is made a comparison with the primary dictionary terms. In the following section this comparison and some additional data about the corpus and the data included into it are presented.

4 Testing

As a result of the Web crawling using as seeds the terms from the first version of the dictionary, as shown in Table 1, an amount of documents relevant to the subject of speech disorders was obtained. Some data about the corpus is now presented in Table 2.

Table 2. Some outline data from the corpus.

Number of initial terms used in the gathering of documents.	15
Number of offline documents added to the corpus.	25
Number of documents obtained at the end of web crawling.	395
Corpus size of plain text in bytes.	3,151,819

After applying the pre-preprocessing described in the previous section and the information retrieval algorithms, the terms shown in Table 3 were found to be the most frequent.

The proposed primary dictionary also included composed terms but in this initial analysis of the corpus just single terms frequency is searched. The original single terms proposed in the primary dictionary can be compared against the single terms found to be the most frequent in the corpus (See Table 4).

Table 3. 15 most frequent terms in corpus.

No.	Term	Frequency
1	Speech	4,036
2	Disorder	2,798
3	Child	2,369
4	Language	1,695
5	Health	1,332
6	Information	1,180
7	Help	963
8	Therapy	949
9	Sound	809
10	Communication	772
11	Research	742
12	Services	695
13	Words	694
14	Development	651
15	Medical	640

Table 4. Comparison of single proposed terms vs single most frequent terms.

No.	Single Term (proposed)	Single term (most frequent)
1	Speech	Speech
2	Disorder	Disorder
3	Dyslalia	Child
4	Dysglosia	Language
5	Dysarthria	Health
6	Dysphemia	Information
7	Therapy	Help

Only the terms *speech* and *disorder* are kept in both lists. Just the first 7 single most frequent terms are used because in the original proposed terms there are just 7 single terms. The rest of the terms that do not appear in the top 7 frequent single terms (*dyslalia, dysglosia, dysarthria, dysphemia* and *therapy*) are listed with their frequency in the Table 5.

Table 5. Frequency in corpus of the rest of proposal single terms.

.

No.	Single Term (proposed)	Frequency in corpus
1	Dyslalia	81
2	Dysglosia	4
3	Dysarthria	437
4	Dysphemia	46
5	Therapy	967

Observing this data from word frequency, not all of the proposed terms in the primary dictionary are equally relevant to the domain of knowledge. Therefore, the web crawler can be fed with the most frequent terms obtained from the corpus and thus, gather more relevant documents.

Another way to complement the corpus is to include synonyms to the original proposed terms. A vast list of terms was found to be included in such list, some of them are listed in the Table 6.

Table 6. Synonyms for some of the original proposed terms.

No.	Original proposed term(s)	Synonyms
1	Speech	Conversation, locution, expression, language, articulation.
2	Disorder	Irregularity, impairment, deficit.
3	Dyslalia	Dysphasia.
4	Dysarthria	Aphasia.
5	Dysphemia, Childhood-onset fluency disorder, Rhythm disorder.	Stammering, stuttering.
6	Speech sound disorder, Communication disorder, Articulation disorder.	Speech impairment, speech impediment, speech defect, delayed speech, speech deficit, speech deficiency, speech disturbance, misarticulation, phonological disorder, phonological delay, phonological impairment, verbal disorder.
7	Therapy.	Treatment, Care.
8	Speech therapy, Logopedic therapy.	Language therapy, Articulation therapy, Speech treatment.
9	Speech development	Speech progress, Speech improvement, Speech maturation, Speech progression.

Applying again the steps of crawling, pre-processing and IR algorithms more documents were added to the corpus and a new list of the most frequent terms is obtained. The Table 7 compares the more frequent terms in the corpus from the last step presented in Table 3 vs. the most frequent terms in the corpus after gathering documents using the synonyms as seeds for crawling.

The 15 most frequent terms obtained after this expansion in the dictionary resulted to be the same as the ones obtained in the previous step non-using synonyms, just varying the order of appearance in the list. Terms as *child* and *language* resulted to be more frequent when synonyms were used as seeds than in the first term frequency list in Table 3.

Table 7. 15 most frequent terms in corpus.

	Primary Dictionary Terms		Extended with Synonyms Dictionary Terms	
No.	**Term**	**Frequency**	**Term**	**Frequency**
1	Speech	4,036	Speech	9125
2	Disorder	2,798	Child	5877
3	Child	2,369	Language	5165
4	Language	1,695	Disorder	4792
5	Health	1,332	Sound	2968
6	Information	1,180	Word	2790
7	Help	963	Health	2786
8	Therapy	949	Information	2697
9	Sound	809	Therapy	2695
10	Communication	772	Help	2081
11	Research	742	Service	1939
12	Service	695	Communication	1687
13	Word	694	Development	1485
14	Development	651	Research	1476
15	Medical	640	Medical	1261

5 Conclusions

The corpus building process proposed for a certain knowledge domain starts with a list of proposed terms followed by a crawling script execution to gather relevant documents. Afterwards, normalizing and IR algorithms were applied to the documents in the corpus in order to include the resulting list of terms into the dictionary; the crawler can be fed again with the new dictionary. Ongoing work consists on the application of word ranking and n-grams algorithms in order to improve the list of terms into the dictionary. Besides, work has been doing in expanding with hyponyms and hyperonyms in the list of terms; this current task allows adding an additional semantic level to the process and it to be able to gather even more relevant documents for the corpus.

Acknowledgements. We would like to thank to the Vicerrectoría de Investigación y Estudios de Posgrado (VIEP) from the Benemérita Universidad Autónoma de Puebla for supporting this work through the project Model of Teaching-Learning Process applying Ontological Engineering.

References

1. NICHCY: Trastornos del habla o lenguaje, 285, 1–4 (2010)
2. Disability in Mexico | Global Disability RightsNow!, http://www.globaldisabilityrightsnow.org/infographics/disability-mexico
3. WHO (World Health Organization): World report on disability 2011. Am. J. Phys. Med. Rehabil. Assoc. Acad. Physiatr (2011). doi:10.1136/ip.2007.018143
4. Loudon, K.: Developing Large Web Applications. O'Reilly Media, California (2010)
5. Robin: What is Corpus?, http://language.worldofcomputing.net/linguistics/introduction/what-is-corpus.html
6. Sharma, S., Ward, E.C., Burns, C., Theodoros, D., Russell, T.: Assessing dysphagia via telerehabilitation: Patient perceptions and satisfaction. Int. J. Speech, Lang. Pathol, 15, 176-183 (2013). doi:10.3109/17549507.2012.689333
7. Schipor, O.A., Pentiuc, S.G., Schipor, M.D.: Automatic assessment of pronunciation quality of children within assisted speech therapy. Elektron, ir Elektrotechnika, 122, 15–18 (2012). doi:10.5755/j01.eee.122.6.1813
8. Martín Ruiz, M.L., Valero Duboy, M.Á., Torcal Loriente, C., Pau de la Cruz, I.: Evaluating a web-based clinical decision support system for language disorders screening in a nursery school. J. Med. Internet Res, 16, e139 (2014). doi:10.2196/jmir.3263
9. Robles-Bykbaev, V., López-Nores, M., Pazos-Arias, J., Quisi-Peralta, D., García-Duque, J.: An Ecosystem of Intelligent ICT Tools for Speech-Language Therapy Based on a Formal Knowledge Model. Stud. Health Technol. Inform., 216, 50–54 (2015). doi:10.3233/978-1-61499-564-7-50
10. Chuchuca-Mendez, F., Robles-Bykbaev, V., Vanegas-Peralta, P., Lucero-Saldana, J., Lopez-Nores, M., Pazos-Arias, J.: An educative environment based on ontologies and e-learning for training on design of speech-language therapy plans for children with disabilities and communication disorders. CACIDI 2016 - Congr. Aergentino Ciencias la Inform. y Desarro. Investigación (2016). doi:10.1109/CACIDI.2016.7785987
11. Ilda, R., Torres, B., López, I.V., Luis, J., Suarez, D.: Aplicación Móvil para la Adquisición de Lenguaje En Niños Con Trastorno De Habla. 40–56 (2016)
12. Kalra, D., Beale, T., Heard, S.: The OpenEHR foundation. http://www.openehr.org/home
13. Haav, H.M.: A Semi-automatic Method to Ontology Design by Using FCA. CLA, 13–24 (2004)
14. Braga, F., Ebecken, N.: A semi-automatic method for extracting a taxonomy for nuclear knowledge using hierarchical document clustering based on concept sets Fabiane Braga. Int. J. Nucl. Knowl. Manag., 6, 155–169 (2013). doi:10.1504/IJNKM.2013.054496
15. Maedche, A., Staab, S.: Semi-automatic engineering of ontologies from text. Proc. 12th Int. Conf. Softw. Eng. Knowl. Eng., 231–239 (2000)
16. Wynne, M.: Developing Linguistic a Guide to Good Practice Corpora (2005)

17. Evans, D.: Corpus building and investigation for the Humanities : An on-line information pack about corpus investigation techniques for the Humanities. Linguistics. 15–16 (2004). doi:10.1109/SocialCom.2010.106
18. Mitchell, R.: Web scraping with Python: collecting data from the modern web. O'Reilly Media, Inc (2015)
19. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders (2013)
20. Python Software Foundation. https://www.python.org/
21. Caio Miyashiro: Text Mining and Natural Language Processing - Preprocessing. http://rstudio-pubs-static.s3.amazonaws.com/67435_ca0769f0dbbb4fc4bda5e4535e21fb54.html
22. Zhu, X.: Common Preprocessing Steps. CS769 Spring 2010 Adv. Nat. Lang. Process, 1–3 (2010)