Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 82–91

Procedia
Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Data science in light of natural language processing: An overview

Imad Zeroual[a]* and Abdelhak Lakhouaja[a]

[a]*Faculty of Sciences, Mohamed First University, Av Med VI BP 717, Oujda 60000, Morocco*

Abstract

The focus of data scientists is essentially divided into three areas: collecting data, analyzing data, and inferring information from data. Each one of these tasks requires special personnel, takes time, and costs money. Yet, the next and the fastidious step is how to turn data into products. Therefore, this field grabs the attention of many research groups in academia as well as industry. In the last decades, data-driven approaches came into existence and gained more popularity because they require much less human effort. Natural Language Processing (NLP) is strongly among the fields influenced by data. The growth of data is behind the performance improvement of most NLP applications such as machine translation and automatic speech recognition. Consequently, many NLP applications are frequently moving from rule-based systems and knowledge-based methods to data-driven approaches. However, collected data that are based on undefined design criteria or on technically unsuitable forms will be useless. Also, they will be neglected if the size is not enough to perform the required analysis and to infer the accurate information. The chief purpose of this overview is to shed some lights on the vital role of data in various fields and give a better understanding of data in light of NLP. Expressly, it describes what happen to data during its life-cycle: building, processing, analyzing, and exploring phases.

Keywords: Data science; Natural language processing; Data driven approches; Corpora; Machine learning

1. Introduction

Recently, a variety of Natural Language Processing (NLP) applications are based on data-driven methods such as neural networks and Hidden Markow Models (HMMs) [1]. Since the progress in most of these methods is driven

* Corresponding author. Tel.: +212618417420.
E-mail address: mr.imadine@gmail.com

10.1016/j.procs.2018.01.101

from data, large and high-quality data become very valuable resources. For instance, some of their beneficial effects have been observed in machine translation [2], word sense disambiguation [3], summarization [4], syntactic annotation [5], named entity recognition [6], among other NLP applications. Over centuries, the primary data have accumulated manually in the form of unstructured repositories of texts called archives. In the early 1980s, the revolution of computer industry leads to a new way of data treatment primarily in term of storage capacity. The new form of an entire population of electronic data are called databases that are designed to facilitate data entry and retrieval. A few years later, the NLP scientists gave a name of corpora to a subset of databases. i.e., a collection of naturally occurring text samples are compiled according to some explicit criteria in order to represent a language [7].

Similarly, various fields such as linguistic, lexicography, and education have been supported by data. This latter, in form of archives, dictionaries, and corpora, are considered as a principal source of evidence for linguistic description and argumentation. Grammarians have always needed sources of evidence as a basis to illustrate grammatical features such as the nature, the structure, and the functions of languages [8]. In lexicography, corpora are exceedingly used to build and make major revision of relevant dictionaries such as the “Dictionary of the Older Scottish Tongue”, the “Middle English Dictionary”, the “Dictionary of Old English”, and the “Oxford English Dictionary” [9]. Alongside the linguistic description and lexicography, data significantly affect a wide range of research activities that have a pedagogical purpose. For instance, learner corpora, collections of first or second language learner data, are generally used to build word frequency lists. These lists are a quick guide and better curricula materials for teaching and learning vocabulary. *Nation* [10] believes that the high-frequency words is important for the learners and need to focus on their learning burden and ensure that the learners will come back to them again. Whereas, some low frequent words may not need to become a part of the learners’ output or the teacher may give some brief attention to them.

The aim of this paper is to outline the main stages in life-cycle of data in NLP, from data design and collection to data processing and analysis. Further, to make this overview equally rich in both theoretical and practical aspects, a survey[a], that covers 100 of well-known and influential corpora, has addressed these stages. Yet, it represents many languages including monolingual corpora (24 languages), bilingual corpora (11 languages), and multilingual corpora (3 to 109 languages). Since the English language was the forerunner in NLP, it is normal that 25% of covered corpora are devoted to English. However, many other languages are catching up, implicitly considering English corpora as a global standard. It worth mention that most covered corpora are publicly available, either free of charge or at an affordable cost, and some of them are available for online search or downloadable. Finally, information regarding the website addresses or *DOI* for all data mentioned in this survey are given in the appendix.

After this introduction, the main content of the paper is structured in section 2. We outline various stages of data life-cycle in light of NLP. Expressly, data sources, format, and corpus design criteria are described. Further, we attempt to present a general view of different data collection methods. Then, data processing and analysis are discussed in detail. Finally, Section 3 contains some concluding remarks.

2. Life-cycle of data in NLP

NLP, also known as computational linguistics, is a subfield of artificial intelligence that aims to learn, understand, recognize, and produce human language content [11]. To get over the limitation of rule-based systems, many research groups move to data-driven methods. However, collecting data is not a trivial undertaking and demands a coherent treatment behind it. For instance, one of the biggest challenges all corpora builders encounter is the lack of public resources as well as copyright. Yet, the quality and quantity of data does matter and should not haphazardly collected. For instance, *Manning* [12] reports that what we need to enhance the part-of-speech tagging and move the accuracy from its current level of about 97.3% to close to 100% is using better training data.

2.1. Data design, sources, and format

A corpus is not haphazard collections of textual material, therefore, a great care must be taken during the data collection [13]. Building corpora usually starts by identifying the appropriate criteria in order to be representative

[a] https://goo.gl/forms/PwxGLe7imVChBPEx2

with respect to the phenomena under investigation [14]. If there are no specific criteria, the corpus should be designed for a general use to suit most NLP applications. In the 90's, primarily with the completion of the British National Corpus [15], basic guidelines in corpus design and compilation have been set by relevant specialists [15–18]. Later, *Sinclair* [19] expanded these guidelines in ten fundamental criteria which are considered as core principles such as representativeness, balance, and homogeneity.

Based on these design criteria, corpora builders should identify the data source genres and size. The selection and finding suitable resources is much more complicated. For example, as stressed by *Lüdeling* [20], the components of general corpora typically are representative of various genres, whilst specialized corpora can be limited to highlight only one genre or a family of genres. Regarding the size and balance, *Lüdeling* claims that it is not always possible to collect data in similar (or even sufficient) quantities for each text category represented in the corpus; this is often the case with historical corpora. However, several sources can be used to build corpora. A corpus, in one hand, may consist of a single book like the one developed by *Baneyx et al*. [21] to build an ontology of pulmonary diseases, and the one used for common-sense knowledge enhanced embeddings to solve pronoun disambiguation problems [22]. On the other hand, corpora are usually developed using a number of books (e.g., Shamela [23]), or editions of a particular newspaper (e.g., [24]). Recently, corpora builders, in particular individuals, use the web to build a very large corpora in a short time and with low cost [25]. However, we must be cautious when drawing data from the web, especially if the aim is to build a balanced corpus in which the language data have to be drawn from a wide range of sources.

Regarding our survey, we proposed 9 well-known sources (books, web, magazines ...), offering the possibility to select multiple choices or adding new sources that are not listed. The results are presented in Table 1.

Table 1. The used sources to build corpora

Sources of corpora	Number of corpora
books	11
books, Official Prints, Newspapers, Magazines	16
Dictionaries	1
Human-generated	4
Newspapers	12
Records	2
Video subtitles	3
Web	22
Web, books, Official Prints, Old Manuscripts, Newspapers, Magazines	22
Web, Human-generated	3
Web, Newspapers, Magazines	4
Total	100

Text encoding or markup is one of the main tasks during data collection. Typically, corpora consist of electronic versions of texts taken from various sources. Therefore, a confusion may arise due to different codes used for markup. Since the 1980s, the Standard Generalized Markup Language (SGML) has become increasingly accepted as a standard way of encoding electronic texts. Using SGML is considered as a basis for data preparation; it facilitates the portability of corpora, enabling them to be reused in different contexts on different equipment, thus saving the cost of repeated typesetting. Since SGML can be complex for some corpora builders and users, an Extensible Markup Language (XML) that was derived from SGML contains a limited feature set to make it simpler to use. Among the different XML standards, the formation of the Text Encoding Initiative (TEI[b]) has to be seen as the first, the most comprehensive, and mature encoding standard for machine-readable texts. Since its launch in 1987, the TEI has become the reference platform for the representation of machine-readable texts [26]. TEI aims to encode as

[b] http://www.tei-c.org/

many possible views (e.g., text components as physical, typographical or linguistic objects) of a text as possible. The TEI P5, which was the current edition at the time this article was first submitted, was released on November 2, 2007.

2.2. Data collection methods

The earlier corpora that occurred before the 1960s are called pre-electronic because they were not computerized. The roots of developing a corpus can be traced back to 1897 when the German linguist *Kaeding* assembled a large corpus that contains 11 million words [27]. Based on this corpus, *Kaeding* could publish the first known word frequency list using word counting. In doing so, he needed the help of over five thousand assistants over a period of years to process the corpus [28, 29]. However, the first corpus to be given that name was the Brown Corpus [30, 31], developed and released in 1964. It consists of texts amounting to over one million tokens after compiling a single year of publication (1961).

Understandably enough, building corpora is time-consuming and often difficult to undertake regarding the size of data and the adopted compilation methods. Through this survey, it was possible to gather information about the methods used to build 100 corpora in relation to their size and the average time consumed during the compilation procedure. The purpose of this study is not to determine the best way to build a corpus. Instead, we would like to know what relevant methods are used and how long they took to complete the procedure. Table 2 exhibits the obtained results.

Table 2. Compilation methods of corpora

Compilation methods	Corpora size (Nb of tokens)	Nb of corpora	Average time consumed
Manual (28%)	<= 100K	5	2 years
	<= 1M	8	3 years 6 months
	<= 10M	7	4 years 6 months
	<= 50M	4	7 years
	<= 100M	1	3 years
	<= 500M	2	3 years
	<= 1Bn	1	7 years
Semi-automated (40%)	<= 100K	1	2 years
	<= 1M	8	3 years
	<= 10M	6	4 years 6 months
	<= 50M	3	5 years
	<= 100M	3	2 years
	<= 500M	10	5 years
	<= 1Bn	3	6 years
	> 1Bn	6	10 years 3 months
Automated (25%)	<= 10M	1	3 years
	<= 50M	1	4 years
	<= 100M	6	5 years 8 months
	<= 500M	5	3 years 10 months
	<= 1Bn	3	2 years 6 months
	> 1Bn	9	3 years

Crowdsourcing (6%)	<= 100K	2	1 years 9 months
	<= 1M	2	7 months
	<= 50M	1	8 years
	<= 1Bn	1	Since 2001
Gamified approach (1%)	<= 100K	1	43 days

Table 2 exhibits the summary of compilation methods mentioned in the survey. We can observe that corpora builders usually rely on three major methods: (1) The first and the oldest method ever, the manual method. (2) The inevitable result after the appearance of computers, the automatic method. (3) This latter is not completely accurate, because the crawled data are often duplicated on the Web and need to be cleaned, filtered, and converted into the right format. Therefore, a manual edition is subsequently performed. This is the semi-automatic method. Concerning the crowdsourcing method, it is catching up. The crowdsourcing was the result of the advancement of digital technology and the Internet in the mid-2000s. It is mainly an online sourcing model in which individuals or organizations use contributions from online communities for problem solving and resources production [32]. One of the main platforms used in crowdsourcing is Amazon's Mechanical Turk[c] (e.g., [33, 34]). The last observed method is the gamified approach or "Gamification". It is a new-coming method to NLP [35], it was first mentioned in 2003 and start to be used in literature in 2010 [36] (e.g., [37]). By definition, gamification is the use of game design elements in non-game contexts to increase users' motivations towards given activities [38]. Though gamified approaches are based on the same strategy as crowdsourcing, in this latter, the participants receive money to increase motivation; whereas gamification incentivizes them by getting entertained.

It can be surmised that producing a corpus with a considerable size and variety requires years of efforts regardless of the used method. In the context of NLP, the first and the most way to collect meaningful and high-quality of data is to hire expert linguists to manually build or annotate corpora; however, it takes time, and costs money. The survey shows that 86% of manually built corpora contain less than 50 million words and it takes more than four years to complete them; whereas, 88% of automatically built corpora with a size that varies from 100 million to billions are completed in less than the same period. Further, the automatic methods marked a considerable progress lately, mainly due to the recent advances in deep learning technologies (e.g., deep neural networks), especially when it comes to dealing with very large data and access to information. The "Intelligent Personal Assistants" like Apple Siri, Google Assistant, Microsoft Cortana, Amazon Echo, etc. are certainly sufficient examples of the significant success achieved using these methods. Generally, the semi-automated methods are a combination of both manual and automated methods, the aim is to raise the quality of data within a feasible/reasonable time and cost. In doing so, at first stage, the data are collected or annotated automatically and later edited by experts. It worth mentioning that the shortest time consumed for building a corpus is achieved by the gamified approach, this method can be promising, especially it combines some key features of the other methods. For example, it is fast as well as an automated method and provides good quality results; yet, it is based on the same strategy as crowdsourcing, but its cost does not scale with data size.

2.3. Data processing and analysis

After the data compilation, a set of general analyses and text processing are essential for NLP scientists. Generally, corpus annotation is a process devoted to assign linguistic features to language data [39]. The annotation implies to use standards, or at least, a well-elaborated set of procedures that is widely accepted and frequently used into common corpus creation and processing tools. Over the last years, significant efforts have been put into developing procedures for automatic annotation and text processing. The aim is to greatly reduce the human intervention and vastly expands the empirical basis.

While some corpora builders attempt to annotate their collected data or processing existing ones, many others still do not bother with processing tasks and they just produce the corpora in raw format. Further, one of the main

[c] https://www.mturk.com

aims being addressed in NLP is developing new forms of annotation and improving the accuracy of automatic annotation. The following tasks are some key features of corpus annotation, they are considered sophisticated as a part of a basic data processing workflow, and their design is more contentious. They can be performed manually by expert linguists to ensure the quality of annotation, or they may be done automatically to allow an extensive and difficult analyses to be carried out. To name a few:

- Lemmatization: The purpose of lemmatization process is to reduce surface words to their canonical form called lemma; this latter is a dictionary lookup form. The lemma relates different word forms that have the same meaning [40]. Further, using lemmatization is found to be efficient, in particular, for information retrieval [41].
- Part of speech (PoS) tagging: It is a basic task in corpus linguistics and an underlying condition to support many subsequent NLP applications. It aims to assign a morpho-syntactical features to each word in a sentence according to the context; also, it allows simple syntactic searches to be performed [42].
- Parsing: The natural successor to PoS tagging is parsing. Basically, it provides a dependency tree as an output. Here, the goal is predicting for each sentence or clause an abstract representation of the grammatical entities and the relations between them. Consequently, the parser assigns a fully labelled syntactic tree or bracketing of constituents to sentences of the corpus [43].

According to the survey, Table 3 presents the different annotation processes performed during corpora building. Note that, many corpora may contain parts of both annotated texts and raw data.

Table 3. Types of corpora annotation

Annotation types	*Number of corpora*
Raw texts	7
Texts with metadata	17
Partially annotated	10
Full annotation	28
Parsed	1
Semantic annotation	1
Parts of Raw texts and Texts with metadata	4
Parts of Raw texts and Full annotation	2
Parts of Texts with metadata and Partially annotated	12
Parts of Texts with metadata and Full annotation	11
Parts of Parsed and Partially annotated texts	2
Parts of Parsed and Full PoS annotated texts	5
Total	100

As seen in Table 3, 40% of corpora contain texts with metadata, and with the same rate, we have corpora with full annotation, which means they contain, at least, lemmatized and PoS tagged data. A small rate (8 corpora) has been totally or partially parsed and only one corpus has been semantically annotated. These results are expected knowing that parsing and sentiment analysis require an advanced level knowledge of linguistic theories and rules. However, automatic sentiment analysis has been increased in recent years due to the rapid data growth in social networks especially twitter.

Regarding the data analysis, they are probably the most important procedures in data life-cycle that could help to solve issues, support theories, observe phenomena, extract valuable information, and produce useful and reliable applications in academic, commercial and public sectors. Next, we briefly state the basis analysis used in NLP:

- Concordance: It aims to search the text and finds every occurrence of each word and displays all the occurrences with its immediate context in a corpus. Further, Concordances can be produced in several formats, but the most usual form is the Key Word in Context (KWIC) concordance [29]. The first concordance, completed in 1230, was that of Hugo of Caro [44], Dominican monk and cardinal (died 1263). It has been said that 500 monks to have been engaged upon its preparation. As the basis of their work they used the text of the Latin Vulgate, the standard Bible of the Middle Ages in Western Europe.

- Word frequency procedure: It aims to produce lists of words and their frequency in the corpus. Moreover, it is feasible to produce word frequency lists using a PoS tagged corpus not merely their orthographic status. For example, Leech et al [45] use the tagged British National Corpus to build the word frequencies list in written and spoken English.
- Collocation statistics: It is a procedure to calculate statistical information about the association, the strength of collocation, and the comparative frequencies of word forms in a corpus or multiple corpora. Unlike the previous analytical procedures, collocation requires a big-sized corpus. Otherwise, the analyst cannot cope with the available data.

In addition, deeper analysis lead to produce NLP applications that help performing sophisticated tasks effectively that no individual or using knowledge-based methods can do better, not to mention that they could require years or decades of efforts. As examples of those applications, Information Retrieval (IR), summarization, and Machine Translation (MT). By applying IR, we can search and retrieve relevant information required among enormous data (e.g., search engines [46]), even if the source documents are written in a language while the user's queries are in another one [47]. Summarization analysis can turn large numbers of documents into shorter versions without changing meaning of the original texts. This is useful for a question answering system, whereas, the summarization generates a sort of summary of the answer, primarily if the question is why or how type question (e.g., [48]). On the other hand, the benefit from online education is still limited by language barriers since the content is usually generated in English. Machine translation can bridge the language gap by providing an initial translation which can be later post-edited by translators (e.g., [49]).

3. Conclusion

In this paper, we have briefly presented distinct stages in the life-cycle of data in NLP. The goal of this overview is to help readers to better understand the data design, collection, annotation, and analysis procedures by taking an experiment from already published corpora of the most prominent projects of resource rich–poor languages. In order to make the survey rich in information, we targeted well-known corpora in the literature and those publicly available; yet, we covered many languages in purpose to ensure that the survey is balanced as much as possible. Finally, this overview is always a subject to make further progress and more detailed descriptions and to a critical understanding of the conceptual, technical, and practical scope of data in light of NLP.

References

1. Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D.: Natural language processing using very large corpora. Springer Science & Business Media (2013).
2. Hu, K., others: Introducing corpus-based translation studies. Springer (2016).
3. Lefever, E., Hoste, V.: Semeval-2013 task 10: Cross-lingual word sense disambiguation. Proc SemEval. 158–166 (2013).
4. Li, L., Forascu, C., El-Haj, M., Giannakopoulos, G.: Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian. Presented at the (2013).
5. Xing, J., Wong, D.F., Chao, L.S., Leal, A.L.V., Schmaltz, M., Lu, C.: Syntaxtree aligner: A web-based parallel tree alignment toolkit. In: Wavelet Analysis and Pattern Recognition (ICWAPR), 2016 International Conference on. pp. 37–42. IEEE (2016).
6. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia. Artif. Intell. 194, 151–175 (2013).
7. Sinclair, J.: Intuition and annotation–the discussion continues. Lang. Comput. 49, 39–59 (2004).
8. Halliday, M., Matthiessen, C.M., Matthiessen, C.: An introduction to functional grammar. Routledge (2014).
9. Milfull, I.: Mutual Illumination: The Dictionary of Old English and the Ongoing Revision of the Oxford English Dictionary (OED3). Florilegium. 26, 235–264 (2009).
10. Nation, I.S.P.: Teaching & learning vocabulary. Boston: Heinle Cengage Learning (2013).
11. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science. 349, 261–266 (2015).
12. Manning, C.D.: Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 171–189. Springer (2011).
13. Leech, G.: Corpora and theories of linguistic performance. Dir. Corpus Linguist. 105–122 (1992).
14. Ball, C.N.: Automated text analysis: Cautionary tales. Lit. Linguist. Comput. 9, 295–302 (1994).
15. Leech, G.: 100 million words of English: the British National Corpus (BNC). Lang. Res. 28, 1–13 (1992).

16. McEnery, T., Wilson, A.: Corpus linguistics. Edinburgh: Edinburgh University Press (1996).
17. Sinclair, J.: Preliminary recommendations on corpus typology. EAGLES Doc. TCWG-CTYPP Available Httpwww Ilc Pi Cnr ItEAGLEScorpustypcorpustyp Html. (1996).
18. Biber, D., Conrad, S., Reppen, R.: Corpus linguistics: Investigating language structure and use. Cambridge University Press (1998).
19. Sinclair, J.: Corpus and text-basic principles. Dev. Linguist. Corpora Guide Good Pract. 1–16 (2005).
20. Lüdeling, A., Kytö, M.: Corpus linguistics: an international handbook. Walter de Gruyter (2008).
21. Baneyx, A., Charlet, J., Jaulent, M.-C.: Building an ontology of pulmonary diseases with natural language processing tools using textual corpora. Int. J. Med. Inf. 76, 208–215 (2007).
22. Liua, Q., Jiangb, H., Linga, Z.-H., Zhuc, X., Weid, S., Hua, Y.: Commonsense Knowledge Enhanced Embeddings for Solving Pronoun Disambiguation Problems in Winograd Schema Challenge. ArXiv Prepr. ArXiv161104146. (2016).
23. Belinkov, Y., Magidow, A., Romanov, M., Shmidman, A., Koppel, M.: Shamela: A Large-Scale Historical Arabic Corpus. ArXiv Prepr. ArXiv161208989. 45 (2016).
24. Maamouri, M., Bies, A., Kulick, S., Gaddeche, F., Mekki, W., Krouna, S., Bouziri, B., Zaghouani, W.: Arabic Treebank: Part 1 v 4.1. (2013).
25. Nakov, P.: Web as a Corpus: Going Beyond the n-gram. In: Russian Summer School in Information Retrieval. pp. 185–228. Springer (2014).
26. Romary, L.: The Text Encoding Initiative: 30 years of accumulated wisdom and its potential for a bright future. In: Language Technologies & Digital Humanities 2016 (2016).
27. Khorsheed, M.S., Alhazmi, K.M., Asiri, A.M.: Developing typewritten Arabic corpus with multi-fonts (TRACOM). In: Proceedings of the International Workshop on Multilingual OCR. p. 16. ACM (2009).
28. Bongers, H.: The History and principles of vocabulary control: as it affects in general and of English in particular. 3. The KLM-List. Wocopi (1947).
29. Kennedy, G.: An introduction to corpus linguistics. Routledge (2014).
30. Francis, W., Kucera, H.: Frequency analysis of English usage. (1982).
31. Hunston, S.: Corpus Linguistics: Historical Development. Encycl. Appl. Linguist. (2013).
32. Brabham, D.C.: Crowdsourcing. Wiley Online Library (2013).
33. El-Haj, M., Kruschwitz, U., Fox, C.: Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. Lang. Resour. Eval. 49, 549–580 (2015).
34. Habernal, I., Gurevych, I.: Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (2016).
35. Zeroual, I., El Kah, A., Lakhouaja, A.: Gamification for Arabic Natural Language Processing: Ideas into Practice. Trans. Mach. Learn. Artif. Intell. 5, (2017).
36. Lund, L., O'Regan, P.: Gamifying Natural Language Acquisition: A quantitative study on Swedish antonyms while examining the effects of consensus driven rewards. (2016).
37. Tiam-Lee, T.J., See, S.: Building a sentiment corpus using a gamified framework. In: Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2014 International Conference on. pp. 1–8. IEEE (2014).
38. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining gamification. In: Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments. pp. 9–15. ACM (2011).
39. Garside, R., Leech, G.N., McEnery, T.: Corpus annotation: linguistic information from computer text corpora. Taylor & Francis (1997).
40. Attia, M., Van Genabith, J.: A jellyfish dictionary for Arabic. In: Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia. pp. 195–212 (2013).
41. Zeroual, I., Lakhouaja, A.: Arabic information retrieval: Stemming or lemmatization? In: 2017 Intelligent Systems and Computer Vision (ISCV). pp. 1–6. IEEE, Fez, Morocco (2017).
42. Zeroual, I., Lakhouaja, A., Belahbib, R.: Towards a standard Part of Speech tagset for the Arabic language. J. King Saud Univ. - Comput. Inf. Sci. 29, 174–181 (2017).
43. Tsarfaty, R., Seddah, D., Kübler, S., Nivre, J.: Parsing morphologically rich languages: Introduction to the special issue. Comput. Linguist. 39, 15–22 (2013).
44. Tribble, C.: Concordancing. Wiley Online Library (2013).
45. Leech, G., Rayson, P., others: Word frequencies in written and spoken English: Based on the British National Corpus. Routledge (2014).
46. Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice. Addison-Wesley Reading (2010).
47. Bhattacharya, P., Goyal, P., Sarkar, S.: Query Translation for Cross-Language Information Retrieval using Multilingual Word Clusters. WSSANLP 2016. 152 (2016).
48. Jayakody, J., Gamlath, T.S.K., Lasantha, W.A.N., Premachandra, K., Nugaliyadde, A., Mallawarachchi, Y.: "Mahoshadha", the Sinhala Tagged Corpus Based Question Answering System. In: Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1. pp. 313–322. Springer (2016).
49. Jansen, D., Alcala, A., Guzman, F.: Amara: A sustainable, global solution for accessibility, powered by communities of volunteers. In: International Conference on Universal Access in Human-Computer Interaction. pp. 401–411. Springer (2014).

Appendix: Corpora of the survey

Corpora	language(s)	Corpus reference (URL or DOI)
Al-Hayat Arabic Corpus	Arabic	http://catalog.elra.info/product_info.php?products_id=632
Alpino Treebank	Dutch	http://odur.let.rug.nl/~vannoord/trees/
Amara	20 languages	http://alt.qcri.org/resources/qedcorpus/
Arab-Acquis	Arabic, English, and French	http://www.aclweb.org/anthology/E/E17/E17-

		2.pdf#page=267
Arabic English Parallel News	Arabic, English	https://catalog.ldc.upenn.edu/LDC2004T18
Arabic Treebank	Arabic	https://catalog.ldc.upenn.edu/LDC2005T20
Aranea Web Corpora	18 languages	http://unesco.uniba.sk/guest/
ARCADE/ROMANSEVAL	English, French, Italian	http://catalog.elra.info/product_info.php?products_id=535
ARCHER	English	http://www.projects.alc.manchester.ac.uk/archer/
arTenTen	Arabic	https://www.sketchengine.co.uk/artenten-corpus/
BAF	French, English	http://rali.iro.umontreal.ca/rali/?q=fr/BAF
BoLC	Italian/English	http://corpora.ficlit.unibo.it/
BOLT	English, Chinese	https://catalog.ldc.upenn.edu/LDC2016T19
Brown	English	http://clu.uni.no/icame/brown/bcm.html
BulTreeBank	Bulgarian	http://www.bultreebank.org/
CELEX2	English, German, Dutch	https://catalog.ldc.upenn.edu/LDC96L14
CELT	Irish, Latin, English, French, Spanish, Italian, Provençal, Dutch, Danish	http://www.ucc.ie/celt/
CETEMPúblico	Portuguese	http://www.linguateca.pt/CETEMPublico/
CINTIL Corpus	Portuguese	http://cintil.ul.pt/
ConVote	English	http://www.cs.cornell.edu/home/llee/data/convote.html
CORGA	Galician	http://corpus.cirp.es/corga/
CORIS	Italian	http://corpora.ficlit.unibo.it/
CORIS/CODIS	Italian	http://corpora.ficlit.unibo.it/
Corpora for eContent professionals	Greek-English, Bulgarian-English, Slovene-English, and Serbian-English	http://dl.acm.org/citation.cfm?id=1706253&CFID=770410841&CFTOKEN=72713156
Corpus "TUITS" IRÓNICOS	Spanish	https://ivanvladimir.github.io/sitio-corpus-ironia/
Corpus del Español	Spanish	http://www.corpusdelespanol.org/hist-gen/
Corpus del Español (Web)	Spanish	http://www.corpusdelespanol.org/web-dial/
Corpus do Português	Portuguese	http://www.corpusdoportugues.org/hist-gen/2008/
Corpus do Português (Web)	Portuguese	http://www.corpusdoportugues.org/web-dial/
Corpus of Spoken Lithuanian	Lithuanian	http://donelaitis.vdu.lt/sakytines-kalbos-tekstynas/
CRATER	English, French, Spanish	http://catalog.elra.info/product_info.php?products_id=636
CREA	Spanish	corpus.rae.es/creanet.html
Croatian National Corpus	Croatian	10.1007/978-1-4020-4068-9_14
Daniel corpus	Chinese, English, Greek, Polish, and Russian	10.13140/2.1.1094.6881
DeReKo	German	http://www1.ids-mannheim.de/kl/projekte/dereko_i.html
deWaC	German	http://wacky.sslmit.unibo.it
DiaCORIS	Italian	http://corpora.ficlit.unibo.it/
EMEA Corpus	22 languages	http://opus.lingfil.uu.se/EMEA.php
English Gigaword	English	https://catalog.ldc.upenn.edu/ldc2011t07
Europarl	21 European languages	http://www.statmt.org/europarl/
Frantext	French	http://www.frantext.fr/
frWaC	French	http://wacky.sslmit.unibo.it
GeFRePaC	German, and French	http://catalog.elra.info/product_info.php?products_id=633
GENIA	English	http://www.nactem.ac.uk/meta-knowledge/download.php
Global English Monitor Corpus	English	http://www.corpus.bham.ac.uk/ccl/global.htm
GUM (Georgtown University Multilayer corpus)	English	https://corpling.uis.georgetown.edu/gum/
Helsinki Corpus	English	http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus
ICE-GB	English	http://www.ucl.ac.uk/english-usage/projects/ice-gb/
International Corpus of English	English	http://www.ucl.ac.uk/english-usage/projects/ice.htm
International Corpus of Learner English - ICLE	25 languages	http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm
INTERSECT	English, French, German	http://arts.brighton.ac.uk/staff/raf-salkie/intersect
itWaC	Italian	http://wacky.sslmit.unibo.it
KACST	Arabic	10.1007/s10579-014-9284-1
KAzakh Dependency Treebank	Kazakh	https://github.com/UniversalDependencies/UD_Kazakh
Kazakh Language Corpus	Kazakh	http://kazcorpus.kz/klcweb/en/
Korean National Corpus	Korean	http://www.sejong.or.kr/
KorpusDK	Danish	http://ordnet.dk/korpusdk_en?set_language=en
KSUCCA)	Arabic	http://ksucorpus.ksu.edu.sa
Lancaster Parsed Corpus	English	http://clu.uni.no/icame/lanpeks.html
Lancaster-Leeds Treebank	English	http://universal.elra.info/product_info.php?cPath=42_43&products_id=437
LASSY	Dutch	http://odur.let.rug.nl/~vannoord/Lassy/

LIVAC	Mandarin Chinese	http://www.livac.org
Malay Concordance Project	Malay	http://mcp.anu.edu.au/
Movie Review Data	English	http://www.cs.cornell.edu/people/pabo/movie%2Dreview%2Ddata/
MultiLing Multilingual Multi-Document Summarization Corpus	Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian, Spanish	http://multiling.iit.demokritos.gr/pages/view/1540/task-mms-multi-document-summarization-data-and-information
MultiUN	English, French, Spanish, Arabic, Russian, Chinese, German	http://www.euromatrixplus.net/multi-un/
Negra	German	http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
NEMLAR Corpus	Arabic	http://www.rdi-eg.com/Projects/nemlar.htm
OntoNotes	English, Chinese, Arabic	https://catalog.ldc.upenn.edu/LDC2013T19
PANACEA	English, French, Greek	http://catalog.elra.info/product_info.php?products_id=1182
Parallela	English-Polish	http://paralela.clarin-pl.eu/
ParTUT	English, French, Italian	https://github.com/msang/partut-repo
Penn Treebank	English	https://catalog.ldc.upenn.edu/ldc99t42
Polarity	English	10.1109/HNICEM.2014.7016215
PTPARL Corpus	Portuguese	http://catalog.elra.info/product_info.php?products_id=1179
Russian collection	Russian	http://corpus.leeds.ac.uk/ruscorpora.html
Russian National Corpus	Russian	http://www.ruscorpora.ru
SIKOR	North Saami, South Saami, Aanaar Saami, Lule Saami, Skolt Saami	http://gtweb.uit.no/korp/#?cqp=%5B%5D&lang=en
Sinica	Chinese	http://ckip.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm
Slovak National Corpus	Slovak	http://korpus.juls.savba.sk/
Syntactic Database for modern Spanish (BDS)	Spanish	http://www.bds.usc.es/
The American National Corpus	English	http://www.anc.org/
The Bank of English	English	http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html
The British National Corpus	English	http://www.natcorp.ox.ac.uk/
The Czech National Corpus	Czech	https://www.korpus.cz/
The Hellenic National Corpus	Greek	http://hnc.ilsp.gr/find.asp
The hungarian gigaword corpus	Hungarian	http://corpus.nytud.hu/mnsz/index_eng.html
The International Corpus of Arabic	Arabic	http://www.bibalex.org/ica
The Lancaster Corpus of Mandarin Chinese	Chinese	http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/default.htm
The National Corpus of Polish	Polish	http://nkjp.pl
The New York Times Annotated Corpus	English	https://catalog.ldc.upenn.edu/LDC2008T19
TIGER Corpus	German	http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html
TIPSTER Complete	English	https://catalog.ldc.upenn.edu/LDC93T3A
TüBa-D/Z treebank	German	http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html
Turkish National Corpus	Turkish	http://www.tnc.org.tr
UKPConvArg	English	https://www.ukp.tu-darmstadt.de/data/argumentation-mining/ukpconvarg1-corpus/
ukWaC	English	http://wacky.sslmit.unibo.it
Wikipedia: Database	More than 270 languages	https://en.wikipedia.org/wiki/Wikipedia:Database_download
WIT3	109 languages	https://wit3.fbk.eu/
WOCHAT	English	http://workshop.colips.org/wochat/