Lexical Data in Multilingual Context: Seeking Cognates Through Syllabic Grouping

Evangelos C. Papakitsos[1]
[1] Dept. of Industrial Design & Production Engineering, University of West Attica, Athens, Greece, papakitsev@uniwa.gr

Abstract. The technique of syllabic grouping, introduced herein, is a computational method of Natural Language Processing for discovering cognates in multilingual corpora and text collections. The lexical cognates are words of the same origin in different languages. This technique is being developed for a project of Digital Humanities, aiming at assisting the decipherment of Linear A script, which is a script of the Bronze Age Aegean area, rendering one or more unknown or lost languages. Nevertheless, it is also suggested that seeking cognates through internet can be very useful for data mining in multilingual environments, especially when the purpose of retrieving information is sentiment analysis, namely the assessment of products and services by their users/customers. Without considering multilingualism, valuable information can be missed. Since the individual opinions collected can be vast, depending on the subject, Big Data technology, which is a component of Industry 4.0, is crucial for sentiment analysis in combination with corpora linguistics. The graphemic matching is described through simple examples that demonstrate its future potentials.

Keywords: *data mining; information retrieval; sentiment analysis; Natural Language Processing; Digital Humanities.*

1. Introduction

The 4th Industrial Revolution is a new episode in industrial evolution, introducing more advanced manners of production and differing in scale, complexity, speed and transformational power, compared to previous revolutions (Schwab & Davis, 2018; Xu et al., 2018). The term "Industry 4.0" is "a collective term for technologies and concepts of value chain organization" (Hermann et al., 2016: 11). This term and the associated

production model include concepts such as the Smart Industry/Factories, Advanced Manufacturing, the Industrial Internet of Things (IoT), the Internet of Services (IoS), the Internet of Media (IoM) and Cyber-Physical Systems (Davies, 2015; Hermann et al., 2015; Lu, 2017; Magruk, 2016; Pereira & Romero, 2017).

In this context, a diversity of interdisciplinary technologies allows the automation, digitization and completion of material production and a manufacturing process, in general (Oesterreich & Teuteberg, 2016), that includes Artificial Intelligence (AI) and Machine Learning, advanced robotics and autonomous transportations, 3D printing, advanced materials, biotechnology and genomics, Cloud Computing and Big Data (Li et al., 2017; World Economic Forum, 2016).

2. Big Data

Big Data is a technological field that deals with the analysis of data sets too complex and large for extracting information with other processing methods, yet being prone to higher false-discovery rate (Breur, 2016). The relevant challenges of data analysis include issues of sources, searching and capturing, visualizing, managing and information privacy, being initially linked to the three key concepts of volume, variety, and velocity (Jain, 2016). Nevertheless, "big" (i.e., volume) is not necessarily the main feature of data, since "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem" (Boyd & Crawford, 2011). Subsequently, certain limitations are encountered in particular scientific fields (Reichman et al., 2011).

Big Data technology is mainly focused on unstructured data, although it deals with structured and semi-structured ones, as well (Dedić & Stanier, 2017). A main problem with Big Data is sampling, namely, whether it is necessary to process the entire set of data or not. The main components of Big Data include (Manyika et al., 2011; Mann & Hilbert, 2020): databases, cloud computing, business intelligence, charts, graphs and other display means, testing techniques, artificial intelligence for development (AI4D), machine learning and Natural Language Processing (NLP).

3. Big Data and NLP

Since natural language is the human tool of communication, a vast amount of information/data that are accessible through internet is found in natural languages. In Linguistics, the large and structured collections of texts are called *corpora* and their study constitutes the subject of *corpus linguistics* (Facchinetti, 2007), aiming at researching language in "real conditions" (i.e., as it is actually used by the speakers, daily). Obviously, the existence of computational methods and tools is of paramount importance for supporting corpus linguistics, being a specific branch of NLP. Therefore, the research interest in combining Big Data with NLP for corpora is intensively expressed (Burkette & Kretzschmar, 2018; Lu, 2020), specifically stating that: "*Corpora are an all-important resource in linguistics, as they constitute the primary source for large-scale examples of language usage. This has been even more evident in recent years, with the increasing availability of texts in digital format leading more and more corpus linguistics toward a "big data" approach. As a consequence, the quantitative methods adopted in the field are becoming more sophisticated and various*" (Marelli, 2019).

Corpora can be either *monolingual*, namely collections of texts written in a single language, or *multilingual*, especially for facilitating the goals of *machine translation*. The monolingual corpora can be vast, especially for the diachronic study of language evolution, occasionally amounting to 1.3 billion words (Renouf, 2018). Such large corpora may result from digitization activities of historical sources, like book collections and newspapers, exhibiting heterogeneous data ("variety") and posing difficult processing challenges (Rupp et al., 2014). Such a processing challenge is *sentiment* or *emotional analysis*, being a branch of NLP that deals with the automatic extraction of the opinion of a person from text, regarding his/her satisfaction or dissatisfaction from using a product or a service. In a large scale, this processing aims at recording the public opinion on various matters (Zhang et al., 2019), a task that also requires data mining techniques, based on machine learning and deep learning, as in the case of Tourism. In this last case, the problem of multilingualism has been revealed. Namely, most of the corpora are written in English, yet the existence of relevant corpora in other languages too makes the processing techniques not universally adaptable (Li et al., 2019). More specifically, an opinion for a certain product or service in nowadays international markets can be written in many languages. Therefore, the sentiment analysis for this particular product or service may be insufficient, when the related big data mining is conducted in a single language. Besides machine

translation, another experimental solution to this problem could be seeking cognates.

4. Lexical Cognates

The lexical cognates are words of common etymological origin (Crystal, 2011) in different languages, that may have similar form and similar (or not necessarily) meaning. For example, the Latin word "corps" remained unchanged in English, French and Dutch, while it became "corpo" in Italian and Portuguese, "cuerpo" in Spanish, "corp" in Romanian, "korps" in German, etc. Seeking cognates is an activity of fundamental importance in historical/comparative linguistics, because this is the tool for discovering the relations between different languages and thus leading to the formation of a linguistic family tree or "phylogeny". Cognates have been utilized for proposing the reconstruction of Proto-Sapiens, the hypothetical language spoken by the homo-sapiens group of people who left Eastern Africa 50-70 thousand years ago, to inhabit the entire planet. According to the theory of "monogenesis", all human languages originate in Proto-Sapiens (Papakitsos & Kenanidis, 2018).

Another application of cognates from Digital Humanities is the deciphering of ancient scripts, rendering unknown or lost ancient languages. Such a successful application achieved the automatic decipherment of Ugaritic, a known but lost language of the Western Semitic family of the 14th century BCE, written in cuneiform consonantal alphabet. The decipherment had been accomplished due to the linguistic similarities of Ugaritic to Hebrew (Papakitsos et al., 2018). Similar is the case with Linear A script that renders one or possibly more unknown ancient and lost languages (Kenanidis & Papakitsos, 2015). It had been used during the Bronze Age (2nd and 3rd millennia BCE), mainly in Minoan Crete but also in other areas of the Aegean Archipelago and beyond. In this respect, a software application is being developed for assisting the study and decipherment of Linear A (Mavridaki et al., 2020). Part of this system is a computational tool that facilitates the discovering of Linear A's cognates in several other contemporary languages, presented next.

5. Seeking Cognates

The main concept of the afore-mentioned computational tool for discovering cognates is the observed existence of a core group of

consonants, common in the various cognates. Looking at the previous example of "corps", it can be observed that the core group of consonants consists of {c, r, p}, common in almost all cognates and in that order of appearance in each word. Depending on the occasion/language, the last consonant {s} is omitted, due to its ending nature that is language specific (as in declension). Therefore, omitting vowels, as well, which are generally more prone to phonetic change (Papakitsos & Kenanidis, 2018), the remaining "core consonantal form" (CCF) of the cognates is CRP. A second stage of "graphemic normalization" is also required to account for the German "korps" that have the CCF of KRP. Namely, the German grapheme {K} has to be matched to the grapheme {C}, based on their phonetic resemblance, to acquire the "normalized" CCF (NCCF) of CRP (the opposite is also possible, depending on the matching convention of the designer/engineer). Thus in searching, the words are substituted by their NCCFs, while the retrieved words with a common NCCF are potential cognates to be further examined.

In the case of the software application for the decipherment of Linear A script, the graphemic normalization is conducted according to the following table of "syllabic grouping" (Table I).

TABLE I. SYLLABIC GROUPING FOR LINEAR A SCRIPT

Original	Matched	Original	Matched
A	A	E - I	I
Ə	A – E	O - U	U
D – T – Δ – Θ	D	W	W
B – F – P – V	B	L – R	L
C - K	C	M	M
G – Q – X – Γ	G	N	N
H – J	Ø	S – Z	S

Table I has been formed according to the observed graphemic/phonetic similarities and correspondences between the words of Linear A script of unknown language(s) and those of Linear B script (Greek). Linear A script can be red but not understood, so far, but both scripts share common/cognate words, mainly denoting anthroponyms, toponyms, names of plants and divinities, etc. On the odd columns of Table 1, the original alphabetic graphemes (transliterated from the actual syllabic script) rendering the corresponding phonetic values are placed in groups, because it has been also observed that certain phonetic values belonging to the same category (i.e., dentals, labials, palatals, etc.) can be interchangeable

between scripts and/or rendered languages. Therefore for example, all dental graphemes {D, T, Δ, Θ}, whenever found in words, are conventionally matched to the first of them {D} (see 2nd column of Table I) and substituted by it at the NCCF of the processed word. Here though, the same syllabic grouping is conducted for vowels too (see top two rows of Table I), thus forming the "normalized core graphemic form" (NCGF) of words. Consequently, the possible cognates of Linear A script to other contemporary languages are examined through the corresponding NCGF of their words, stored in a database.

6. Conclusion

The research in Digital Humanities may result in useful tools for NLP, regarding sentiment analysis through big data mining, in multilingual environments. It has been exemplified herein that in multilingual corpora the technique of syllabic grouping could assist the discovery of cognates in different languages, thus not missing potentially valuable information for the assessment of products and services. The syllabic grouping of keyword cognates is a computational task much simpler than machine translation. Therefore, the former task may precede the latter one on keywords, in order to examine whether machine translation during big data mining is worthy of conducting or not. In this manner, time and volume could potentially be saved, resulting in more manageable situations of information retrieval. The technique of syllabic grouping is just being implemented and tested for assisting the decipherment of Linear A script; it remains to be seen in the future whether it will be proved useful or not elsewhere too.

Acknowledgement. The author expresses his thankfulness to Assist. Prof. A. Lengeris for his comments on issues of syllabic grouping.

References

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. In Symposium on the Dynamics of the Internet and Society, Social Science Research Network: A Decade in Internet Time. DOI: 10.2139/ssrn.1926431

Breur, T. (2016). Statistical Power Analysis and the contemporary "crisis" in social sciences". Journal of Marketing Analytics, 4(2–3), 61–65.

Burkette, A., & Kretzschmar, W., Jr. (2018). Big Data: Using a Corpus. In Exploring Linguistic Science: Language Use, Complexity, and Interaction (pp. 201-210). Cambridge: Cambridge University Press.

Crystal, D. (ed.) (2011). Cognate. Blackwell Publishing: A Dictionary of Linguistics and Phonetics (6th ed., p. 104). ISBN 978-1-4443-5675-5.

Davies, R. (2015). Industry 4.0 Digitalisation for productivity and growth. European Parliamentary Research Service, 2015, 1-10.

Dedić, N., Stanier, C. (2017). Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. In Lecture Notes in Business Information Processing, 285, Innovations in Enterprise Information Systems Management and Engineering (pp. 114–122). Berlin; Heidelberg: Springer.

Facchinetti, R. (2007). Theoretical Description and Practical Applications of Linguistic Corpora. Verona: QuiEdit. ISBN 978-88-89480-37-3.

Hermann, M., Pentek, T., & Otto, B. (2016). Design Principles for Industrie 4.0 Scenarios. In the 49th Hawaii International Conference on System Sciences (HICSS), pp. 3928-3937. DOI: 10.1109/HICSS.2016.488

Hermann, M., Pentek, T., & Otto, B. (2015). Design principles for industry 4.0 scenarios: A literature review. Dortmund: Technische Universität Dortmund.

Jain A. (2016). The 5 V's of big data. IBM: Watson Health Perspectives.

Kenanidis, I.K., & Papakitsos, E.C. (2015). A Comparative Linguistic Study about the Sumerian Influence on the Creation of the Aegean Scripts. Scholars Journal of Arts, Humanities and Social Sciences, 3(1E), 332-346.

Li, G., Hou, Y., & Wu, A. (2017). Fourth Industrial Revolution: technological drivers, impacts and coping methods. Chinese Geographical Science, 27(4), 626-637.

Li, Q., Li, S., Zhang, S., Hu, J., & Hu, J. (2019). A Review of Text Corpus-Based Tourism Big Data Mining. Applied Sciences, 9, 3300.

Lu, T. (2020). Analysis on Linguistics Research Directions in the Age of Big Data. Journal of Physics: Conference Series, 1606, 012008.

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1-10.

Magruk, A. (2016). Uncertainty in the sphere of the industry 4.0–potential areas to research. Business, Management and Education, 14(2), 275-291.

Mann, S., & Hilbert, M. (2020). AI4D: Artificial Intelligence for Development. International Journal of Communication, 14, 21.

Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxburgh, C., Hung Byers, A. (2011). Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Marelli, M. (2019). Quantitative Methods in Morphology: Corpora and Other "Big Data" Approaches. Oxford Research Encyclopedia of Linguistics.

Mavridaki, A., Galiotou, E., & Papakitsos, E.C. (2020). Designing a Software Application for the Multilingual Processing of the Linear A Script. In Proceedings of the 24th Pan-Hellenic Conference on Informatics (PCI 2020), Athens, Greece. ACM International Conference Proceedings Series.

Oesterreich, T.D., & Teuteberg, F. (2016). Understanding the implications of digitisation and automation in the context of Industry 4.0: A triangulation approach and elements of a research agenda for the construction industry. Computers in Industry, 83, 121-139.

Papakitsos E.C., Kenanidis I.K. (2018). Going to the Root: Paving the Way to Reconstruct the Language of Homo-Sapiens. International Linguistics Research, 1(2): 1-16.

Papakitsos, E.C., Kontogianni, A., Papamichail, C., & Kenanidis, I.K. (2018). An Application of Software Engineering for Reading Linear-B Script. International Journal of Applied Science, 1(2), 58-67.

Pereira, A.C., & Romero, F. (2017). A review of the meanings and the implications of the Industry 4.0 concept. Procedia Manufacturing, 13, 1206-1214.

Reichman, O.J., Jones, M.B., Schildhauer, M.P. (2011). Challenges and opportunities of open data in ecology. Science, 331(6018), 703–705.

Renouf, A. (2018). Big Data: Opportunities and Challenges for English Corpus Linguistics. In C. Suhr, T. Nevalainen, & I. Taavitsainen (Eds.), From Data to Evidence in English Language Research (pp. 27–65). Leiden: Brill.

Rupp, C. J., Rayson, P., Gregory, I., Hardie, A., Joulain, A., & Hartmann, D. (2014). Dealing with heterogeneous big data when geoparsing historical corpora. In 2014 IEEE International Conference on Big Data (Big Data), pp. 80-83, DOI: 10.1109/BigData.2014.7004457

Schwab, K., & Davis, N. (2018). Shaping the future of the fourth industrial revolution. Redfern, NSW: Currency.

World Economic Forum (2016). The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution. Geneva: Global Challenge Insight Report.

Xu, M., David, J.M., & Kim, S.H. (2018). The fourth industrial revolution: opportunities and challenges. International Journal of Financial Research, 9(2), 90-95.

Zhang, W., Zhu, Y. & Wang, J. (2019). An intelligent textual corpus big data computing approach for lexicons construction and sentiment classification of

public emergency events. Multimedia Tools and Applications, 78, 30159–30174.