INTRODUCTION

The past is mediated predominantly by textual material. Iconography, audio recordings, artifacts, human memory have their significance, but even they are described, discussed, referred to or transcribed with the written word. This is particularly relevant for the five hundred-year-old Gutenberg era, especially as its end (marked by the ascent of the electronic media) starts to fade below the horizon of living memory.

Even if the past is not distant, human memory (be it individual, be it collective) might be patchy. Let us consider a linguistic example: the verb *ogarniać*, in a number of senses, has been rising in popularity in colloquial Polish. Polish speakers might remember the time when it was not frequent and might be able to testify that they could hear it all around them in a given year, but few of them are likely to be able to pinpoint the year when the fad started. In order to do this, looking through 2000s digital-born texts, or 1990s printed material if needed, might be the only option.

Texts from the past might be combed for various types of, very broadly speaking, *linguistic units*, though the common denominator might not be discernible across discipline, as in linguistics these are morphemes, words (a word as such or in a specific sense), collocations, idiomatic expressions, proverbs, in folkloristics: motifs, folk tales, jokes, contemporary legends, chain letters, in literary studies: tropes, literary references, in history: references to people, countries, cities, inventions, social phenomena, institutions, events, rumours, in economics: price listings, exchange rate tables, company reports. In all cases, a linguistic unit is to be attested with excerpts, or preferably *photo excerpts*

(fragments clipped from original paper publications scanned and made available in an electronic form).

Such needs are as old as the invention and spread of the alphabet and humanity had its ways to address them in an "analogue" manner: through paper dictionaries, motif indexes, bibliographies, library catalogues. With the advent of the electronic computer, these crude measures can be substituted with direct access to the large masses of texts.

The aim of this book is to lay out the theory and practice of *mining* such linguistic units in diachronic collections of both digitised and digital-born publications. The focus is on Polish, though, in principle, the ideas could be transplanted to other languages preserved in print and available online. The program could be given as twelve directives comprising theoretical assumptions, technical measures as well as the guidelines for organisation of collective research. The directives are a generalisation of the twelve directions proposed in the context of contemporary legends in folkloristics[1] and, simultaneously, build on the theory of linguochronologisation developed by Piotr Wierzchoń.[2]

D1. All excerpts and photo excerpts related to the linguistic unit in question should be stored in a common database available as a Web application for all members of a given research community. Every researcher could add new linguistic units and excerpts. There should be a possibility for researchers to express minority or dissenting views (e.g. whether an excerpt really exemplifies a given linguistic unit or not). The database could be initiated with texts gathered in a traditional manner (by scanning physical clippings excerpted manually in linguistics, by entering tales collected from informants in folkloristics, etc.). In particular, the database should be merged with existing chronologisation data, i.e. linguistic units accompanied by temporal metadata: Słownik Bibliograficzny Języka Polskiego,[3] Depozytorium Leksykalne Języka Polskiego,[4] Narodowy Fotokorpus Języka Polskiego.[5]

D2. A *metasearch engine*, i.e. a search engine that aggregates and organises results returned by other search engines (including duplicate removal), should be made available to researchers within the Web service proposed in (D1). With a metasearch engine, laborious entering of the same queries

1. F. Graliński. "Folklorystyka 2.0". In: *NETLOR. Wiedza cyfrowych tubylców*. Ed. by P. Grochowski. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2013, pp. 119–130.
2. Wierzchoń, "Torując drogę teorii lingwochronologizacji".
3. J. Wawrzyńczyk. *Słownik bibliograficzny języka polskiego: wersja przedelektroniczna. T.1, A-Ć*. Warszawa: Instytut Informacji Naukowej i Studiów Bibliologicznych Uniwersytetu Warszawskiego, 2000.
4. P. Wierzchoń. *Depozytorium leksykalne języka polskiego. Nowe fotomateriały z lat 1901–2010*. Vol. Tom I. Warszawa: Bel Studio, 2010.
5. http://nfjp.pl retrieved 1 February 2019.

(related to a given linguistic unit) into multiple search engines is no longer required. Issues specific to a given language, e.g. inflection for Polish, could be taken into account within a metasearch engine, even if the upstream search engines have their limitations (e.g. inflection is not considered in phrase searches in Google search engine, which is a problem when looking for occurrences of a nominal phrase – a metasearch engine could generate all the possible forms of a phrase).

D3. The texts indexed by general Web search engines (Google, Bing, etc.) should be made available in the metasearch engine (if technically possible). Every two days humanity creates as much information as it did up to 2003 (estimated at 5 exabytes[6]) and each new webpage, blog post or tweet might bring new linguistic units or their new variants. It should be even feasible to discover them the moment they appear, in real time.

D4. The mass of texts grows not just into the future, but also into the *past*, as more and more efforts are made to preserve cultural heritage all over the world and more and more print material – old newspapers, books, documents, posters, photographs and even school certificates[7] or train tickets[8] – is being digitised and made available online. This makes it possible to discover *old* linguistic units. In particular, vast amounts of information are becoming available due to continuous and systematic digitisation initiatives to create *digital libraries* (collections of content that are both digitised and organised).[9] All texts available from Polish digital libraries should be aggregated and made accessible for researchers (linguists, historians, genealogists, etc.) and searchable in the full-text manner (this might be limited to publications from the late 18th century and later, for which the quality of OCR output is good enough).

D5. Historical texts are not limited to printed publications, as *webpages* are becoming archival materials. There exist a number of initiatives to archive World Wide Web, the largest of which is Internet Archive, a US-based non-profit institution.[10] The websites archived there, including Polish webpages[11] and now-defunct websites are publicly available from the Wayback Machine service.[12] Unfor-

6. P. Lyman and H. Varian. *How much information 2003?* http://www.sims.berkeley.edu/how-much-info-2003. 2004.

7. See e.g. http://www.sbc.org.pl/Content/51267

8. See e.g. http://dlibra.karta.org.pl/catl/Content/13495

9. See also: F. Graliński. "Polish digital libraries as a text corpus". In: *Proceedings of 6th Language & Technology Conference*. Ed. by Z. Vetulani and H. Uszkoreit. Poznań: Fundacja Uniwersytetu im. Adama Mickiewicza, 2013, pp. 509–513.

10. https://archive.org retrieved 1 February 2019.

11. See, for instance, onet.pl as captured by Internet Archive in February 1997: https://web.archive.org/web/19970220173653/http://www.onet.pl.

12. https://web.archive.org retrieved 1 February 2019.

tunately, the search options are still rather limited there. Texts extracted from Polish webpages should be made available within the metasearch engine. This would enable researchers to discover linguistic units in old webpages, especially webpages (personal websites, message boards) abounding in vernacular expression and being rich sources of informal language and contemporary folklore.

D6. Older historical texts can be found on the Internet outside institutional digital libraries, through various local or grassroots initiatives.[13] For instance, the website of the Niepokalanów monastery includes the collected writings of St. Maximilian Mary Kolbe (e.g. personal letters);[14] the writings are available as plain texts (rather than scans), so photo excerpts cannot be made out of them, but it is still a valuable, fully-searchable source of data for linguists and historians. Web resources of this type should be catalogued, their metadata (including temporal metadata, i.e. creation or publication dates) should be extracted and their texts should be *indexed* (*indexing* is a technical term for creating data structures enabling rapid search and retrieval). One of the first Web resources of this type was Virtual Library of Polish Literature founded in 2001.[15] Some initiatives to provide Polish historical publications as plain texts (i.e. without scans) has reached significant size, for instance the ChronoPress chronological corpus comprises manually curated texts from Polish newspapers (1945–1954).[16] Such material is eclipsed by much larger quantities of OCR-ed publications accessible from digital libraries, but still can complement them (e.g. for the purposes of evaluating OCR-cleaning procedures).

D7. Articles posted in Polish Usenet newsgroups should also be indexed. Usenet is a world-wide distributed discussion system, popular in 1995–2005, later in decline, as people shifted to Internet message boards, blogs and social media. Contrary to some social media today, all the Usenet discussions were carried on openly and are an important source of language data for the turn of the millennium, especially as informal linguistic units are concerned. The order of magnitude of compressed plain text (along with metadata) to be extracted from Polish Usenet newsgroups is gigabytes.

13. M. Wilkowski. *Wprowadzenie do historii cyfrowej*. Gdańsk: Instytut Kultury Miejskiej, 2013, p. 52.

14. More on this in Section 1.3.2.

15. https://literat.ug.edu.pl/books.htm retrieved 1 February 2019.

16. http://chronopress.clarin-pl.eu retrieved 2 February 2019; see also: A. T. Pawlowski. "Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish." In: *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, 2016, pp. 311–313.

D8. Search services offered by Polish daily newspapers (e.g. *Gazeta Wyborcza*,[17] *Rzeczpospolita*[18]) should be made available within the metasearch engine. (Even though full access to the texts is behind a paywall, their search engines are freely available.)

D9. After resources in (D3–D8) are aggregated under a metasearch engine, the next step is to automate search processes. A computer should be "trained" to identify a given linguistic unit (a word in a specific sense, a reference to a specific person, a type of contemporary legend) after a few examples are tagged manually by a researcher. Advanced computational techniques based on statistics and machine learning should be applied to this aim. Some results have already been achieved as far as identification of Polish urban legends is concerned,[19] [20] which is harder than clustering, for instance, news, as proper names are mostly useless for urban legends (the same story wanders from city to city).

D10. All (photo) excerpts with linguistic units found (either manually or automatically) should be stored in the system so that a researcher would not need to save them manually. Excerpts should be clipped automatically.

D11. Members of the international research community should be given access to the system so that equivalent linguistic units in languages other than Polish could be entered (e.g. stories of the same type in other languages in folkloristics). Furthermore, the system should be able to identify equivalent foreign linguistic units using advanced techniques such as machine translation[21] or cross-language information retrieval (i.e. searching for texts in a language different from the one in which the query was formulated).

D12. Everything that can be automated, must be automated. Humans should focus on what is not automatable, i.e. analytical work.

This book reports on an attempt to implement, at least to some extent, this program: (D1, D2, D4, D6, D10) and partially (D9) were realised.

The book is organised into four parts. In the first part, it is discussed how to gather and manage large masses of diachronic texts. The second part presents linguistic and technical issues

17. http://www.archiwum.wyborcza.pl retrieved 1 February 2019.

18. http://archiwum.rp.pl retrieved 1 February 2019.

19. R. Grundkiewicz. "Automatyczne wyszukiwanie i grupowanie krótkich tekstów narracyjnych zamieszczonych w Internecie". PhD thesis. Uniwersytet im. Adama Mickiewicza w Poznaniu, 2011.

20. R. Grundkiewicz and F. Graliński. "How to Distinguish a Kidney Theft from a Death Car? Experiments in Clustering Urban-Legend Texts". In: *Proceedings of the RANLP 2011 Workshop on Information Extraction and Knowledge Acquisition*. Hissar, Bulgaria: Association for Computational Linguistics, Sept. 2011, pp. 29–36.

21. Machine translation does not have to be perfect to be able to match Polish linguistic units with foreign material.

related to the full-text search in historical publications, whereas advanced text modelling techniques are discussed in the third part. Finally, the fourth part showcases specific linguistic as well as folkloristic applications of the methods described in the first three parts.