Anna Dąbrowska / Tadeusz Piotrowski

Information Computer Technologies and the Polish language

Streszczenie

Technologie komputerowe i informatyczne stosowane są w Polsce przede wszystkim w celu korzystania z Internetu, komunikowania się, zdobywania informacji. Ich wykorzystanie bardziej specjalistyczne dopiero ostatnio staje się powszechniejsze dzięki zwiększającym się umiejętnościom użytkowników komputerów, aczkolwiek cały czas pozostawia to wiele do życzenia. Zasoby internetowe związane z tekstami w języku polskim to przede wszystkim korpusy tekstów współczesnych (np. Narodowy korpus języka polskiego tworzony dla Wielkiego słownika języka polskiego i niewielkie korpusy tworzone dla projektów naukowych, np. korpus błędów popełnianych przez cudzoziemców uczących sie języka polskiego) i historycznych (np. Cyfrowa Biblioteka Druków Ulotnych). W instytutach Polskiej Akademii Nauk czy też na Politechnice Wrocławskiej prowadzone są prace nad komputerową analizą współczesnego języka polskiego (analiza morfosyntaktyczna i semantyczna); istnieją też narzędzia kontrolujące pisownię, gramatykę i styl tekstu (Language Tool).

W ramach zachowania dziedzictwa narodowego zostało zdygitalizowanych wiele dawnych tekstów, w tym słowników i encyklopedii, dokumentów i dzieł literackich, historycznych opracowań naukowych (np. gramatyki języka polskiego). Polska Biblioteka Cyfrowa i Biblioteka Narodowa „Polona" oraz kilkanaście bibliotek regionalnych udostępniają w formie cyfrowej zbiory dzieł dotyczących Polski. Istnieją i coraz bardziej rozwijają się biblioteki cyfrowe, również rynek książek elektronicznych (e-booków). W zasobach internetowych nie ma, niestety, wielu informacji dotyczących samego języka polskiego (np. gramatyk, ćwiczeń gramatycznych i stylistycznych, innych pomocy dydaktycznych). Najobszerniejszym opracowaniem jest amatorska gramatyka G. Jagodzińskiego. Niewiele jest też możliwości nauki języka polskiego jako obcego.

Artykuł wspomina też w dużym skrócie wpływ narzędzi elektronicznych (komputera, telefonu komórkowego) na język polski, np. unikanie diakrytyków, wprowadzanie skrótów (wiąże się to również z wpływem języka angielskiego na współczesną polszczyznę).

Information computer technologies are commonly used in Poland for routine tasks, such as use of the Internet, and, as a result, for easy interpersonal communication or in using information services, such as library catalogues. More advanced technologies were earlier developed by computer specialist centres in Poland, but it is only fairly recently that their work has been known to the public at large, especially to linguists and other academics and non-academics dealing with Polish. This paper is a very brief survey of those technologies that can be potentially used by lay persons, who use resources and tools that are commonly known, i.e. in Poland predominantly on the Windows platform.

It also has to be said that the traditional divide between the humanities and the sciences is extremely sharp in Poland. As a result most arts students and scholars use computer technologies at a very basic level and are simply not aware of new possibilities opened up by them. This also results from the scope of studies, narrowed down to traditional "philological" courses. It follows from this that more advanced uses of text processing tools, for example, meet formidable psychological barriers, which can be seen in the facts that the establishment of large corpora of Polish started very late and that they are still not widely used. This also means that there are few studies that are based on corpus research.

Without any doubt the most important current resource, with accompanying suites of tools, is the National Corpus of Polish, which now contains more than 1.5 billion running (text) words (as on August 8, 2011). The corpus is being made for the Wielki Słownik Języka Polskiego (a large non-commercial academic dictionary of Polish), under preparation, and originally was a compilation of existing large corpora of several institutions, with the Instytut Języka Polskiego PAN as a coordinator. These institutions are Instytut Podstaw Informatyki PAN (Institute of Computer Sciences at Polish Academy of Sciences), Wydawnictwo Naukowe PWN (Polish Scientific Publishers PWN) and Zakład Językoznawstwa Komputerowego i Korpusowego, Instytut Filologii Angielskiej Uniwersytetu Łódzkiego (Department of Computational and Corpus Linguistics, English Department, the University of Łódź) (http://nkjp.pl/). At present the corpus can be described as opportunistic, though it aims at being in part at least representative. It contains a subcorpus of speech. The corpus is marked-up for inflection, by two schemes. The site has a working demo of the whole corpus, and, interestingly, it can be accessed from two subsites, each of which has texts with a different mark-up and different tools. One can use not only concordancing facilities, but also look up collocations or time profiles.

There are more specialist, small corpora worked on at various universities. One example is the corpus at Wrocław University (Polish Department), which collects errors made by students of Polish as a foreign language. It has 14,400 errors, recorded in typical sentential contexts. They are used in descriptions of those difficulties of Polish in grammar or in the lexicon that students have most problems with. This corpus is supplemented on a regular basis.

There is not much interest in corpora of historical periods of Polish. The most important one is the corpus of Old Polish (texts that can be dated from before 1500), which, unfortunately, contains only a selection of the most significant texts (Biblioteka Zabytków Polskiego Piśmiennictwa Średniowiecznego Instytucie Języka Polskiego PAN w Krakowie, 2006; http://kupujmy.eu/product_info.php?products_id=63). This is available on DVD-ROM, which contains both graphical images of the texts and the transliterated text, with the relevant editorial description. The transliteration can be also downloaded, free of charge, from the website of the Institute of Polish, Polish Academy of Sciences (http://www.ijp-pan.krakow.pl/index2.php?strona=korpus_tekst_star). Another historical corpus is Digital Library of Polish and Poland-Related News Pamphlets from the 16th to the 18th Century (http://cbdu.id.uw.edu.pl/cgi/set_lang?langid=en&fromurl=http://cbdu.id.uw.edu.pl/), which contains DjVu images (without the text layer) of texts in various languages.

As mentioned above, natural language processing has been studied in several centres in Poland. Linguists are perhaps better acquainted with The Linguistic Engineering Group in Warsaw (headed by Adam Przepiórkowski), which is part of the Department of Artificial Intelligence at the Institute of Computer Science, Polish Academy of Sciences and with The WrocUT Language Technology Group G4.19 (headed by Piasecki), at the Department of Artificial Intelligence, Institute of Informatics, Wroclaw University of Technology.

Both groups work on development of tools to process Polish and, specifically, the Warsaw group was one of the first to develop a large corpus of Polish, free of charge, and

applications to process Polish texts. A number of them have been released on the GNU General Public License, including a powerful application to work with corpora, Poliqarp (http://poliqarp.sourceforge.net/), or a Polish morphosyntactic tagger, TaKIPI. Unfortunately, these cannot be used "off the shelf" as they often require programming skills. Therefore it is unlikely that they will exert any influence on the "traditional" linguistic community. The Wrocław group works on similar applications. They are initially concerned with developing a Polish WordNet, a network of lexical-semantic relations, and an electronic thesaurus with a structure modelled on that of the Princeton WordNet. In contrast to other WordNets, it is based on extraction of items related semantically from a corpus, and is also aimed at automatic language processing (http://plwordnet.pwr.wroc.pl/main/?lang=en).

The most widespread application used to work with text is Microsoft Word, which is most often used like a typewriter, even in very complex projects for which it is clearly not suitable (for example to compile dictionaries). Word is available with the suite of commonly used programs to help produce Polish text: a spelling checker, a thesaurus and a grammar and style checker. Macintosh computers also have a corresponding suite (when Microsoft Office is not used). While it is deplorable that Microsoft, thanks to questionable selling practices, managed to oust from the market native Polish applications, with their own correction tools, OpenOffice is gaining its share of the Polish market and it has its own standard linguistic tools for Polish. OpenOffice can be used with a powerful correction tool for several languages, which was developed in Poland: LanguageTool (www.languagetool.org/). It can be used for Polish, English, German, Russian, etc., and, apart from a spelling module, it is a style and grammar checker, exceeding the quality the (Microsoft) product for Polish. Moreover, it can be used for strictly linguistic tasks, such as the morphological tagging of a corpus, and it was in fact used for tagging the one billion word National Corpus of Polish.

The most widely known – and used – linguistic publication is a dictionary. At present most commercial general dictionaries, both monolingual and bilingual, are available in digital versions. This includes the largest contemporary dictionaries of Polish, such as *Uniwersalny słownik języka polskiego PWN*,[1] *Multimedialny* (*Inny słownik języka polskiego*[2]), and bilingual dictionaries, for example *Oxford-PWN* or *Nowy słownik Fundacji Kościuszkowskiej*[3] for English and Polish. Most often these dictionaries are offered both as standalone applications on DVD-ROMs and as paid up services on web pages (for a monolingual dictionary cf. http://usjp.pwn.pl/, for bilingual ones: http://oxford.pwn.pl/ or http://www.kosciuszkowski.org/). The largest publisher of dictionaries in Poland, PWN, also offers a number of simplified monolingual dictionaries online at http://sjp.pwn.pl/ http://so.pwn.pl. Simple monolingual and bilingual dictionaries can also be accessed by users of mobile networks.

It is quite interesting that apparently publishers do not want to issue digital versions of specialist dictionaries, for example those of synonyms, idioms, or collocations. One specialist dictionary of interest, especially to non-native speakers of Polish is *Słownik gra-*

[1] Ed. by Stanisław Dubisz, Warszawa: PWN 2003; CD-ROM.
[2] Ed. by Mirosław Bańko, Warszawa: PWN 2000; CD-ROM.
[3] Ed. by Jacek Fisiak et al., Kraków: Universitas 2008.

matyczny języka polskiego[4] on CD-ROM, which includes inflection patterns and all inflectional forms of most words in Polish (with about 244,000 entries it is perhaps the largest list of contemporary Polish lexemes; cf. http://sgjp.pl/).

Dictionaries also play an important role in the projects of digital preservation of the national heritage and at present most significant historic Polish dictionaries and encyclopaedias can be accessed on Internet pages, at least in part. Because there is little cooperation between various libraries and individuals interested in digital storage, some of these are available in several copies, for example the monumental encyclopedia of geographical entities, *Słownik geograficzny Królestwa Polskiego i innych krajów słowiańskich*,[5] or the first monolingual dictionary of Polish, which is in fact also a multilingual translation dictionary, in six volumes, *Słownik języka polskiego* by Linde[6] (the above dictionaries at http://poliqarp.wbl.klf.uw.edu.pl/). Other notable nineteenth century dictionaries are: *Słownik warszawski*[7] (http://poliqarp.wbl.klf.uw.edu.pl/), still the largest Polish dictionary and *Słownik wileński*[8] (http://swil.zozlak.org/?prototyp=). Most of the pages at the sites referred to do not contain only images, but their text can be searched and concordances generated.

In addition, contemporary historical dictionaries, of various periods, are either being transferred to the digital medium, such as the huge unfinished dictionary of the 16th century, *Słownik polszczyzny XVI wieku*[9] (http://poliqarp.wbl.klf.uw.edu.pl/slownik-polszczyzny-xvi-wieku/ or http://kpbc.umk.pl/publication/17781), or are compiled from scratch in the digital medium (the one of the 17th and the 18th centuries, at http://xvii-wiek.ijp-pan.krakow.pl/pan_klient/). On the other hand, the most significant dictionary of the 20th century, edited by Witold Doroszewski,[10] was available for some time only on CD-ROM as low-quality images of pages because of unsolved copyright issues. Apparently there is not much interest in the digitization of other metalinguistic books, such as grammars, with the exception of historic nineteenth century texts, for example *Gramatyka* by Onufry Kopczyński (http://babel.hathitrust.org/cgi/pt?id=nyp.33433016467197) or those by Józef Muczkowski (http://babel.hathitrust.org/cgi/pt?id=uc1.b86970).

Dictionaries can be found at digital libraries, which in general collect significant texts in Polish or relating to Poland. After a period of uncontrolled development, work on digitization is now coordinated on a national scale by librarians. There are at least several dozen digital libraries in Poland, two national in scope (Polska Biblioteka Internetowa, www.pbi.edu.pl/index.html and Cyfrowa Biblioteka Narodowa "Polona", www.polona.pl/dlibra), nine regional ones, most often hosted by university libraries, and numerous local ones (cf. www.bg.umcs.lublin.pl/nowa/literat.php). A number of projects are under way.

---

[4] Saloni, Z./Gruszczyński, W./Woliński, M./Wołosz, R. (2007): *Słownik gramatyczny języka polskiego,* Warszawa: Wiedza Powszechna.

[5] T. 1-15, pod red. Filipa Sulimirskiego, Bronisława Chlebowskiego, Władysława Walewskiego, Warszawa (1880-1914).

[6] Samuel Bogumił Linde (1807-1814): *Słownik języka polskiego*, t. 1-6. Warszawa.

[7] A widely used name for: Karłowicz, J./Kryński, A./Niedźwiedzki, W. (1900-1927): *Słownik języka polskiego*, t. 1-8. Warszawa.

[8] A widely used name for: *Słownik języka polskiego* (1861), t. 1-2. Wilno: Wyd. Maurycy Orgelbrand.

[9] *Słownik polszczyzny XVI wieku*, t. I-XXXIV (1966-). Wrocław/Warszawa: Ossolineum and Instytut Badań Literackich PAN.

[10] Doroszewski, W. (ed.) (1958-1969): *Słownik języka polskiego*, v. 1-12. Warszawa: PWN.

There is an agreed de facto standard of software in Poland, so called dLibra Digital Library Framework, developed since 1999 locally in Poznań, and based on DjVu technology, which uses graphical images, with an optional text layer. A number of libraries use this platform (cf. Digital Library of Wielkopolska, www.wbc.poznan.pl/dlibra). While there are more and more texts available in digital libraries (the Wielkopolska Library has more than 100,000 items), their quality, especially metadata, leaves much to be desired. Thanks to the interest in mobile access to digital texts, several companies now offer their own ebooks and sell dedicated readers (cf. www.eclicto.pl/ or www.empik.com/ebooki).

At present the Internet is very often the first and the most important source of information about a topic, and it is interesting what one can find there about the Polish language. Unfortunately, there is not so much informational content (if one does not take into account general reference works, such as Wikipedia), most web pages are in Polish and have been created by non-specialists, who often have a prescriptivist attitude to their language. In what follows we will disregard pages created by non-native Polish specialists, such as the one by Oscar Swan, the University of Pittsburgh (http://polish.slavic.pitt.edu/), which contains, for example, a course on Polish, a grammar and a bilingual dictionary, which can be downloaded free of charge.

A traditional grammar, quite detailed, in Polish and English, was written by Grzegorz Jagodziński, a biologist; it can be found at http://grzegorj.private.pl/gram/pl/gram00.html. It contains also reviews and various thoughts on linguistic matters. While there are practically no serious descriptions of Polish, there are a number of linguistic counselling services (poradnie językowe), which answer questions about "correct usage" on the web. They evolved from phone services and are usually run by universities or publishers (a tentative list of those services can be found at www.poradniajezykowa.us.edu.pl/index.php?action=inne_por).

What can usually be found on the web, in Polish, is various descriptions of particular points or areas in the language,[11] usually aimed at students from primary or secondary schools; they are often called *cribs*.[12] One can also find notes for exams in descriptive grammar (http://gramatyka.wordpress.com/home/). As usual, their value is uneven.

There is a survey of Polish rural dialects on the web, produced by specialists in the field, with descriptions of dialects, examples of texts and recordings (www.gwarypolskie.uw.edu.pl/index.php?option=com_frontpage&Itemid=1). These pages are in Polish only.

A person who would like to learn Polish using web resources will not find rich resources. Courses that can be found are usually basic ones, and were predominantly created in international projects (e.g., www.oneness.vu.lt, www.slavic-net.org), some were created by non-specialists, for example the person mentioned above, Grzegorz Jagodziński, offers a very traditional course of Polish for beginners,[13] based on memorizing metalinguistic description. There are no courses for more advanced learners.

[11] For example: www.sciaga.pl/tekst/55419-56-imieslowy_sciaga, www.sciaga.pl.

[12] For example: http://www.bryk.pl/teksty/gimnazjum/j%C4%99zyk_polski/gramatyka/24908-podstawowe_wiadomo%C5%9Bci_z_gramatyki_j%C4%99zyka_polskiego.html, http://ruczjak.webpark.pl/gramatyka.htm.

[13] http://grzegorj.w.interia.pl/kurs/index.html.

On the Web page of the School of Polish Language and Culture, University of Silesia, one can find[14] some programs for teaching or testing skills in Polish: *Grampol* (for intermediate learners) and *Frazpol* (for teaching idiomatic expressions). To teach or practise spelling one can use applications for teaching native Polish children.[15]

While any change in a language is very slow, there are some tendencies that can be seen in the use of Polish under the influence of the computer, the Internet and mobile phones. These tendencies are seen above all in informal texts, and informal uses of the language. On formal occasions Polish text will have the features of pre-computer texts. One conspicuous tendency, in part inherited from the early stages in the development of both hardware and software, is the omission of Polish diacritics (for example, earlier text messages, i.e., SMSs, that used diacritics were far longer than those without them). However, at present there are no technical problems with diacritics and this tendency can be attributed to a certain fashion: diacritic-less text is probably considered to be more colloquial, more trendy, etc. While obviously loss of diacritics can lead to misunderstanding of the message, usually context, linguistic or non-linguistic, is sufficient for disambiguation.

Another tendency in economizing the written form occurs when sequences of characters are shortened – this is perhaps also motivated by more complex morphological and phonological changes in Polish – by a widespread use of clipping, in which typically only the final syllables are elided, and the resulting structure is disyllabic (e.g. *spokojnie >spoko*). Earlier clipping in Polish was not used very often. It can be found not only in single words but also in phrases (typically in highly conventional ones, e.g. *na razie> nara*). Also proper names can be affected by this process. The occurrence of other devices can be attributed to the influence of English, rather than computers, such as use of acronyms (often in English), emoticons, etc.

One important factor, which can exert a powerful influence on Polish speakers, results from various forms of social communication in the Internet, in which non-standard Polish is used. In writing Poles can use their linguistic creativity, which has been stifled in the predominantly highly prescriptive models used in schooling. They can also see that in fact they can quite efficiently communicate using those forms, in contrast to what prescriptivists usually say. Finally, non-standard forms can spread widely and quickly over the networks, thus perhaps accelerating language change.

References

Bień, J.S. (2006): Kilka przykładów dygitalizacji słowników. In: *Poradnik Językowy* 8, 5-63.

Godzic, W. (2000): Język w Internecie: czy piszemy to, co myślimy? In: Bralczyk, J./Mosiołek-Kłosińska, K. (ed.): *Język w mediach masowych*. Warszawa: Upowszechnianie Nauki – Oświata „UN-O“, 178-185.

Grzenia, J. (2007): *Komunikacja językowa w Internecie*. Warszawa: Wydawnictwo Naukowe PWN.

[14] http://www.sjikp.us.edu.pl.

[15] Np. http://www.dyktanda.net/; http://www.tylkoprogramy.pl/ortografia.php.

Piotrowski, T. (2005): Digitization of Polish historic(al) dictionaries. In: *Преглед НЦД* (*Review of the National Center for Digitization*) 6, 4, 95-102. www.komunikacija.org.yu/komunikacija/casopisi/ncd/6/index_e.

Piotrowski, T./Szafran, K. (2005): The dictionary of Polish of the 16th c. and the computer: from paper to (structured) file. In: Kieler, F./Kiss, G./Pajs, J. (eds.): *Computational lexicography*. Budapest: Hungarian Academy of Science, 171-180.

Przepiórkowski, A./Górski, R.L./Lewandowska-Tomaszczyk, B./Łaziński, M. (2009): Narodowy Korpus Języka Polskiego. In: *Biuletyn PTJ*, LXV, 47-55.