META-NET White Paper Series

Languages in the European Information Society

– Polish –

Early Release Edition

META-FORUM 2011

27-28 June 2011

Budapest, Hungary

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under contracts T4ME (Grant Agreement 249119), CESAR (Grant Agreement 271022), METANET4U (Grant Agreement 270893) and META-NORD (Grant Agreement 270899).

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses educators, journalists, politicians, language communities and others.

The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ for each language. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a European Commission Network of Excellence, has conducted an analysis of current language resources and technologies. This analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are many significant research gaps for each language. A more detailed, expert analysis and assessment of the current situation will help maximise the impact of additional research and minimize any risks.

META-NET consists of 47 research centres from 31 countries that are working with stakeholders from commercial businesses, government agencies, industry, research organisations, software companies, technology providers and European universities. Together, they are creating a common technology vision while developing a strategic research agenda that shows how language technology applications can address any research gaps by 2020.

META-NET
DFKI Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin
Germany

office@meta-net.eu
http://www.meta-net.eu

Author

Dr. Marcin Miłkowski, Institute of Computer Science, Polish Academy of Sciences

Acknowledgements

The publisher is grateful to the authors of the German white paper for permission to reproduce materials from their paper.

Table of Contents

Executive Summary

Many European languages run the risk of becoming victims of the digital age because they are underrepresented and under-resourced online. Huge regional market opportunities remain untapped today because of language barriers. If we do not take action now, many European citizens will become socially and economically disadvantaged because they speak their native language.

Innovative, language technology (LT) is an intermediary that will enable European citizens to participate in an egalitarian, inclusive and economically successful knowledge and information society. Multilingual language technology will be a gateway for instantaneous, cheap and effortless communication and interaction across language boundaries.

Today, language services are primarily offered by commercial providers from the US. Google Translate, a free service, is just one example. The recent success of Watson, an IBM computer system that won an episode of the Jeopardy game show against human candidates, illustrates the immense potential of language technology. As Europeans, we have to ask ourselves several urgent questions:

- Should our communications and knowledge infrastructure be dependent upon monopolistic companies?
- Can we truly rely on language-related services that can be immediately switched off by others?
- Are we actively competing in the global market for research and development in language technology?
- Are third parties from other continents willing to address our translation problems and other issues that relate to European multilingualism?
- Can our European cultural background help shape the knowledge society by offering better, more secure, more precise, more innovative and more robust high-quality technology?

This whitepaper for the Polish language demonstrates that a lively language technology industry and research environment exists. Although there is a number of technologies and resources for Polish, they are less numerous than the ones available for English. The technologies and resources are also of poorer quality.

According to the assessment presented in this report, immediate action must be taken before any breakthroughs for the Polish language can be achieved.

META-NET contributes to building a strong, multilingual European digital information space. By realising this goal, a multicultural union of nations can prosper and become a role model for peaceful and egalitarian international cooperation. If this goal cannot be achieved, Europe will have to choose between sacrificing its cultural identities or suffering economic defeat.

A Risk for Our Languages and a Challenge for Language Technology

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in digitised and network communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

We are currently witnessing a digital revolution that is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication and knowledge exchange were accomplished by efforts like Luther's translation of the Bible into common language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled an exchange across languages;
- the creation of journalistic and bibliographic guidelines assured the quality and availability of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology helped to automate and facilitate many of the processes:

- desktop publishing software replaces typewriting and typesetting;
- Microsoft PowerPoint replaces overhead projector transparencies;
- e-mail sends and receives documents faster than a fax machine;
- Skype makes Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- search engines provide keyword-based access to web pages;
- online services like Google Translate produce quick and approximate translations;
- social media platforms facilitate collaboration and information sharing.

Although such tools and applications are helpful, they currently cannot sufficiently implement a sustainable, multilingual European information society, a modern and inclusive society where information and goods can flow freely.

Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European foreign ministers speak in their native language. We might want a platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

It is clear that communication needs have a different quality as compared to a few years ago. In a global economy and information space, more languages, speakers and content confront us and require us to quickly interact with new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter and YouTube) is only the tip of the iceberg.

A global economy and information space confronts us with more languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognize that it is in a language we do not understand. According to a recent report requested by the European Commission, 57% of Internet users in Europe purchase goods and services in languages that are not their native language. (English is the most common foreign language followed by French, German and Spanish.) 55% of users read content in a foreign language while only 35% use another language to write e-mails or post comments on the web.[1] A few years ago, English might have been the lingua franca of the web—the vast majority of content on the web was in English—but the situation has now drastically changed. The amount of online content in other languages (particularly Asian and Arabic languages) has exploded.

An ubiquitous digital divide that is caused by language borders has surprisingly not gained much attention in the public discourse; yet, it raises a very pressing question, “Which European languages will thrive and persist in the networked information and knowledge society?”

Which European languages will thrive and persist in the networked information and knowledge society?

Our Languages at Risk

The printing press contributed to an invaluable exchange of information in Europe, but it also led to the extinction of many European languages. Regional and minority languages were rarely printed. As a result, many languages like Cornish or Dalmatian were often limited to oral forms of transmission, which limited their continued adoption, spread and use.

The approximately 60 languages of Europe are one of its richest and most important cultural assets. Europe’s multitude of languages is also a vital part of its social success.[2] While popular languages like English or Spanish will certainly maintain their presence in the emerging digital society and market, many European languages could be cut off from digital communications and become irrelevant for the Internet society. Such developments would certainly be unwelcome. On the one hand, a strategic opportunity would be lost that would weaken Europe’s global standing. On the other hand, such developments would conflict with the goal of equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society.[3]

The wide variety of languages in Europe is one of its most important cultural assets and an essential part of Europe’s success.

Language Technology is a Key Enabling Technology

In the past, investment efforts have focused on language education and translation. For example, according to some estimates, the European market for translation, interpretation, software localisation and website globalisation was € 8.4 billion in 2008 and was expected to grow by 10% per annum.[4] Yet, this existing capacity is not enough to satisfy current and future needs.

Language technology is a key enabling technology that can protect and foster European languages. Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates regardless of language barriers or computer skills. Language technology already assists everyday tasks, such as writing e-mails, conducting an online search or booking a flight. We benefit from language technology when we:

Language technology helps people collaborate, conduct business, share knowledge and participate in social and political debates across different languages.

- find information with an Internet search engine;
- check spelling and grammar in a word processor;
- view product recommendations at an online shop;
- hear the verbal instructions of a navigation system;
- translate web pages with an online service.

The language technologies detailed in this paper are an essential part of innovative future applications. Language technology is typically an enabling technology within a larger application framework like a navigation system or a search engine. These white papers focus on the readiness of core technologies for each language.

In the near future, we need language technology for all European languages that is available, affordable and tightly integrated within larger software environments. An interactive, multimedia and multilingual user experience is not possible without language technology.

Opportunities for Language Technology

Language technology can make automatic translation, content production, information processing and knowledge management possible for all European languages. Language technology can also further the development of intuitive language-based interfaces for household electronics, machinery, vehicles, computers and robots. Although many prototypes already exist, commercial and industrial applications are still in the early stages of development. Recent achievements in research and development have created a genuine window of opportunity. For example, machine translation (MT) already delivers a reasonable amount of accuracy within specific domains, and experimental applications provide multilingual information and knowledge management as well as content production in many European languages.

Language applications, voice-based user interfaces and dialogue systems are traditionally found in highly specialised domains, and they often exhibit limited performance. One active field of research is the use of language technology for rescue operations in disaster areas. In such high-risk environments, translation accuracy can be a matter of life or death. The same reasoning applies to the use of language technology in the health care industry. Intelligent robots with cross-lingual language capabilities have the potential to save lives.

There are huge market opportunities in the education and entertainment industries for the integration of language technologies in games, edutainment offerings, simulation environments or training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just a few more examples where language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a further need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology represents a tremendous opportunity for the European Union that makes both economic and cultural sense. Multilingualism in Europe has become the rule. European businesses, organisations and schools are also multinational and diverse. Citizens want to communicate across the language borders that still exist in the European Common Market. Language technology can help overcome such remaining barriers while supporting the free and open use of language. Furthermore, innovative, multilingual language technology for European can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

Multilingualism is the rule, not an exception.

Challenges Facing Language Technology

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. We cannot wait ten or twenty years for significant improvements to be made that can further communication and productivity in our multilingual environment.

The current pace of technological progress is too slow to arrive at substantial software products within the next ten to twenty years.

Language technologies with broad use, such as the spelling and grammar features in word processors, are typically monolingual, and they are only available for a handful of languages. Applications for multilingual communication require a certain level of sophistication. Machine translation and online services like Google Translate or Bing Translator are excellent at creating a good approximation of a document's contents. But such online services and professional MT applications are fraught with various difficulties when highly accurate and complete translations are required. There are many well-known examples of funny sounding mistranslations, for example, literal translations of the names *Bush* or *Kohl*, that illustrate the challenges language technology must still face.

Language Acquisition

To illustrate how computers handle language and why language acquisition is a very difficult task, we take a brief look at the way humans acquire first and second languages, and then we sketch how machine translation systems work—there's a reason why the field of language technology is closely linked to the field of artificial intelligence.

Humans acquire language skills in two different ways. First, a baby learns a language by listening to the interaction between speakers of the language. Exposure to concrete, linguistic examples by language users, such as parents, siblings and other family members, helps babies from the age of about two or so produce their first words and short phrases. This is only possible because of a special genetic disposition humans have for learning languages.

Humans acquire language skills in two different ways: learning examples and learning the underlying language rules.

Learning a second language usually requires much more effort when a child is not immersed in a language community of native speakers. At school age, foreign languages are usually acquired by learning their grammatical structure, vocabulary and orthography from books and educational materials that describe linguistic knowledge in terms of abstract rules, tables and example texts. Learning a foreign language takes a lot of time and effort, and it gets more difficult with age.

The two main types of language technology systems acquire language capabilities in a similar manner as humans. Statistical approaches obtain linguistic knowledge from vast collections of concrete example texts in a single language or in so-called parallel texts that are available in two or more languages. Machine learning algorithms model some kind of language faculty that can derive patterns of how words, short phrases and complete sentences are correctly used in a single language or translated from one language to another. The sheer number of sentences that statistical approaches require is huge. Performance quality increases as the number of analyzed texts increases. It is not uncommon to train such systems on texts that comprise millions of sentences. This is one of the reasons why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, available online information, and translation services such as Google Search and Google Translate rely on a statistical (data-driven) approach.

The two main types of language technology systems acquire language in a similar manner as humans.

Rule-based systems are the second major type of language technology. Experts from linguistics, computational linguistics and computer science encode grammatical analysis (translation rules) and compile vocabulary lists (lexicons). The establishment of a rule-based system is very time consuming and labour intensive. Rule-based systems also require highly specialised experts. Some of the leading rule-based machine translation systems have been under constant development for more than twenty years. The advantage of rule-based systems is that the experts can more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. Due to financial constraints, rule-based language technology is only feasible for major languages.

Polish in the European Information Society

General Facts

With about 40-48 million native speakers, Polish is the most spoken West Slavic language around the world. It is the official language of Poland.[5] The auxiliary minority languages that can be used in legal contexts are: German in the west areas of Poland (22 communes using it as auxiliary language), and Belarusian in the east (3 communes), Kashubian (2 communes) and Lithuanian (1 commune).[6]

In Poland, it is the common spoken and written language and the native language of the vast majority of the population, and it is quite homogenous, while the differences between its dialects (góralski, from Podhale region, Silesian in Silesia, and the dialect of Poznań) are fairly small. The minority nationals are the Germans (according to the minority speakers: 300,000 to 400,000), Belarusians (250,000 to 300,000), Ukrainians (300,000), Lithuanians (30,000), Russians (20,000), Slovaks (15,000), Czechs (3,000), Jews (5,000) and Armenians (1,500). The ethnic minorities are the Ruthenians (50,000), the Roma (20,000), the Tatars (2,000) and the Karaites (150). The only regional group recognised is the Kashubians (250,000 to 300,000), with their own regional language. In total, 1,200,000 people belong to regional and national minorities, even though the latest census statistics from 2002 on ethnicity and nationality only note 417,000, including the Germans (147,000), Belarusians (48,000), Ukrainians (34,000), Slovaks (2,000). The strongest concentrations of minority nationals can be found in the provinces of Warmia-Masuria, Podlachia and Opole.

Particularities of the Polish Language

Polish exhibits some specific characteristics, which contribute to the richness of the language but are challenges for computational processing of natural language.

Some of these characteristics allow the speakers to express ideas in a wide variety of ways. First, word order is relatively free in Polish sentences, and it is used to stress the importance of information rather than simply follow from the rules of grammar. Consider e.g. the English sentence

The woman gave the man an apple.

In English, there are two more ways to express the same idea, namely:

The woman gave an apple to the man.

An apple was given to the man by the woman.

In Polish, there exist at least nine possible ways (even though some of them are less likely to be used):

Kobieta dała mężczyźnie jabłko.

Kobieta mężczyźnie dała jabłko.

Kobieta mężczyźnie jabłko dała.

Jabłko mężczyźnie dała kobieta.

Jabłko kobieta dała mężczyźnie.

Jabłko dała kobieta mężczyźnie.

Mężczyźnie jabłko dała kobieta.

Mężczyźnie jabłko kobieta dała.

Mężczyźnie kobieta dała jabłko.

The meaning of these sentences, though grammatically equivalent, varies, as the word order shows which part is the new information in the sentence, and what is already known.

Second, Polish is relatively morphologically rich, which means that for roughly 180 thousand base forms of words, almost 4 million inflected word forms exist. The inflection paradigms are complex, and even their exact number is a matter of a dispute (single exceptions might be thought to create a new paradigm). Even native speakers have problems with properly inflecting many words, and most speakers of Polish as a second language never completely master the complexities of the inflectional system.

Third, many computer applications assume either English or Western-European alphabets, and that may lead to problems with typing Polish diacritical characters (“ą”, “ę” etc.). Historically, it was one of the biggest problems to get international software to work with Polish, and there were numerous ways to encode these characters. Even now, there are at least three popular code pages used for Polish: Unicode (mostly UTF-8), ISO standard and Windows code page (1250). For this reason, older data might easily be corrupted with incorrect encoding. Restoring the proper diacritical characters is not a trivial problem: there are many words that could be created by changing some of the characters to Polish diacritics (for example, “glosy” may be a correct singular genitive form of “glosa” or plural nominative of “głos” that has “l” instead of “ł”).

Other specific characteristics of Polish that make automatic processing of language difficult are the tendency to use comparably long and nested sentences. In addition, the lack of articles makes detection of noun phrases relatively hard, as the only way to detect them is to rely on morphological information (case, number, gender), which is far from unambiguous.

Recent developments

English language is one of the biggest sources of neologisms and calques, in particular in science and technology, and it exerts a considerable influence on contemporary Polish. The number of words loaned from English into Polish is however much lower than in Dutch or German because of the problems with inflecting some words and differences in pronunciation systems. In early 1990s, just after the major political changes, companies used brands that sounded “English like”. Even a grocery shop could bear an English signboard “Your shop”. Today, such a name would be considered ridiculous by a much larger group of speakers. Yet, calques from English, such as “dokładnie” (exactly) or “wydawać się być” (seem to be), are numerous and popular.

Another influence of English is the appearance of more direct forms of address, especially in advertising. While in the past, using the Polish pronoun “ty” (“you” singular) would have been considered rude, it is quite popular these days. Arguably, this influence stems from incorrect, non-professional translations from English, yet it is a stable phenomenon. Similarly, Polish speakers are now more likely to follow English punctuation patterns, especially a comma after introductory phrase, which is, according to traditional Polish punctuation rules, incorrect. Even some typographical characters (such as “&”), never used in Polish before, are borrowed from English.

The previous sources of linguistic influence, such as Soviet propaganda and doublespeak, are now of almost no importance. The official register is now more connected with the bureaucracy of the EU. Though one can find a new tendency towards creating word compounds such as “speckomisja” (special committee) or “Rywingate”, which remind of the older Soviet newspeak compounds, the development seems to be independent from the historical influence of Russian and is connected with English instead, though acronyms are a considerably rarer phenomenon in Polish than in English.

One of the current developments in Polish is that feminine forms for professions are nowadays more frequently used, though they still remain somewhat outside of the official register. Political correctness is also visible in new forms used to refer to foreign nationals, and immigrants from Africa (the word “murzyn” [negro], previously considered neutral, is now all but banned in newspapers).

One of the traditional complaints about the development of Polish is the proliferation of obscene language and brutality in colloquial speech. It must be stressed, however, that these claims are not based on corpus-based historical research.

Some of the traditional inflection patterns seem to undergo a process of simplification (for example, speakers are more likely to say “mieliłem” than “mełłem”, which would be the standard form), and some of the forms become almost extinct in everyday speech. This is especially true of the vocative case in colloquial Polish. Some of the words are also specially simplified to humorous effect in colloquial speech, e.g., instead of the full word “impreza” (party) one could hear “impra”, “klima” instead of “klimatyzacja” (aircondition), or “kolo” instead of kolega (mate). This said, inflection patterns are still highly complex and no simple trend towards simplifying them is discernible.

Language cultivation in Poland

The legal status of the Polish language within the territory of the Republic of Poland is defined more precisely by the Law of 7 October 1999 on the Polish language, with its subsequent amendments (in 2000, 2003, 2004 and 2005). The regulations of this Act relate to “the protection of the Polish language” and to the use thereof in the pursuit of public tasks, in trade and in the fulfilment of labour-law regulations within the territory of the Republic of Poland. The protection of the Polish language shall consist especially: in concern for the correct usage of language and the establishment of conditions for the proper development of language as an instrument of human communication; in counteracting the vulgarisation of the language; in the dissemination of knowledge about language and its role in culture; in the promotion of respect for regional language variations and dialects and the prevention of their extinction; in the promotion of the Polish language in the world and in support for the teaching of Polish in Poland and abroad.

Entities carrying out public tasks within the territory of the Republic of Poland transact all official business and submit statements of intent in the Polish language, unless specific regulations state otherwise. This applies to statements of intent, applications and other forms submitted to official organs of the state (Article 5).

As regards commercial activities, according to Article 7, in commercial dealings involving the participation of consumers and in the fulfilment of labour-law regulations, the Polish language is to be used if the consumer or employee have their place of domicile in the territory of the Republic of Poland at the time an agreement was concluded and this agreement is to be carried out in the territory of the Republic of Poland. In commercial dealings not involving the participation of consumers, the Polish language is to be used only if this trade is carried out by the entities subordinated to the organs of the State or to the regional public authorities.

The obligation to use the Polish language in commercial dealings involving the participation of consumers applies especially to the names of goods, services, offers, guarantee terms, invoices, bills and receipts as well as warnings and consumer information required by separate regulations, operating instructions and information about the properties of goods and services. The obligation to use the Polish language in information on the properties of goods and services also applies to advertising.

Foreign-language descriptions of goods and services as well as foreign-language offers, warnings and consumer information required on the basis of other regulations must be simultaneously made available in a Polish-language version. Descriptions in the Polish language are not required as regards warnings and consumer information, user manuals and information on the properties of goods if they are expressed in universally comprehensible graphic form; if the graphic form is accompanied by a description, it should be drawn up in the Polish language.

Action may be taken against individuals or businesses that do not respect these requirements. Fines are chargeable for infractions.

Supervision of the use of the Polish language is exercised within the scope of their tasks by the President of the Office of Competition and Consumer Protection, the Trade Inspectorate and the district (municipal) consumer spokesman and the State Labour Inspectorate.

According to Article 8, documents, including in particular agreements involving consumers and labour-law agreements, are to be drawn up in the Polish language. The documents may be simultaneously drawn up in one or more language versions. Unless parties decide otherwise, the basis for the interpretation of such documents is their Polish-language version. A job agreement or other document arising out of labour-law regulations, as well as an agreement to which a consumer is a party, may be drawn up in a foreign language at the request of a job-performing party or consumer who is a citizen of a European Union member-state other than the Republic of Poland and has previously been informed of the right to draw up an agreement in the Polish language. A job agreement or other labour-law document may be drawn up in a foreign language at the request of the job-performing party who is not a Polish citizen, and also in the event the employer is a citizen of a European Union member-state or is based in that state.

Polish is the language of teaching, examinations and diploma dissertations in public and non-public schools of all types, in higher state and non-state schools, in educational establishments and other educational institutions, unless specific regulations state otherwise (a growing number of universities offer programmes in English, though). According to the ordinance of the Minister of National Education and Sport of 15 October 2003, the State Commission for the Certification of Proficiency in Polish as a Foreign Language is the supreme body which supervises administration of examinations and issues certificates of the Polish language proficiency at three levels. The foreigner or the Polish citizen, residing abroad, receives an official certificate of proficiency in Polish after examination before the state examination commission.

The regulations of the Act on the Polish Language do not pertain to proper names, foreign daily newspapers, periodicals, books or computer programs with the exception of their description or instructions; the teaching and research activities of schools of higher education, schools and classes with a foreign language of instruction or bilingual instruction, foreign-language teachers' colleges and also the teaching of other subjects, if it is in accordance with detailed regulations; scientific and cultural creativity; customarily used scientific and technical terminology; trade-marks, brand names and indications of the origin of goods and services; norms introduced in the original language in accordance with standardisation regulations.

The authoritative institution that expresses opinions and gives advice on issues concerning the use of the Polish language is the Polish Language Council (Rada Języka Polskiego), acting as a committee of the Polish Academy of Sciences. Every second year, it presents a report on the protection of the Polish language to the Parliament of the Republic of Poland.

The Council, upon a motion by the minister in charge of culture and the protection of national heritage, the minister in charge of education and training and the minister in charge of higher education, the President of the Office of Competition and Consumer Protection, the Chief Inspector of the Trade Inspectorate or the President of the Polish Academy of Sciences, or at its own initiative, expresses by means of a resolution its opinion on the use of the Polish language in public activities and in trade within the territory of the Republic of Poland involving consumers or the execution in the Republic of Poland of labour-law regulations, and establishes the principles of the Polish language's orthography and punctuation.

Learned societies, associations of authors and higher schools (i.e. tertiary schools or universities) may refer any issues on the use of the Polish language to the Council. In the event of significant doubts arising in its official business concerning Polish-language usage any state or local government authorities may seek the opinion of the Council. Producers, importers and distributors of goods or services which do not have an appropriate name in Polish may request the Council for an opinion concerning appropriate terms for the said goods or services.

Besides the Polish Language Council, some other national institutions are engaged (according to their statutes) in the cultivation, protection and/or promotion of the Polish language.

The law which amended the law on the Polish language (11 April 2003) created a legal foundation for officially certifying knowledge of Polish as a foreign language. Two depositions from the Ministry of National Education and Sports dated 15 October 2003 allow foreign nationals to receive certificates confirming their level of knowledge of the Polish language. There are three levels: elementary, intermediate, and advanced. In some countries, the Polish language is prized as giving access to Polish universities and the Polish job market.

The PISA study, conducted in 2009, shows that Polish students performed well above OECD average with respect to reading literacy (a second European country after Finland), the eight best place.[7] This means that language teaching is successful in Poland, though it is arguable that relative linguistic homogeneity contributes to this result.

Polish on the Internet

In spring 2011, almost 55% of the Poles were Internet users.[8] 72% of them said they were online every day. Among young people, the proportion of users is even higher. The existence of an active Polish-speaking web community is also mirrored by the fact that the Polish Wikipedia, with around 800 thousand entries, is one of the largest Wikipedias after English, German, and French (not counting automatically translated versions such as the Thai Wikipedia), and is comparable to the Italian version.[9]

With about 2 million Internet domains in May 2011[10], Poland's top-level country domain .pl is one of the top country extensions in the world.[11] This dominant Internet presence suggests that there is a considerable amount of Polish language data available on the web. In addition, some multi-lingual resources like the online dictionary mash-up ling.pl[12] are freely available.

For language technology, the growing importance of the Internet is important for two reasons. On the one hand, the large amount of digitally available language data represents a rich source for analysing the usage of natural language, in particular by collecting statistical information. On the other hand, the Internet offers a wide range of application areas for language technology.

The most commonly used web application is certainly web search, which involves the automatic processing of language on multiple levels, as we will see in more detail the second part of this paper. It involves sophisticated language technology, differing for each language. For Polish, this comprises matching “ę” and “e” to match texts written without diacritic characters; moreover, all inflected versions of query words should also be found to enhance the search (so not only „wziąłem”, but also „wziąć”, „wzięłam”, „wziąłby”, „wziąwszy…”). But internet users and providers of web content can also profit from language technology in less obvious ways, e.g., if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, comparatively little usable language technology is built compared to the anticipated need. This might be due to the complexity of the Polish language and the number of technologies involved in typical LT applications.

In the next chapter, we will present an introduction to language technology and its core application areas as well as an evaluation of the current situation of LT support for Polish.

Selected Further Reading

Jerzy Bralczyk, *Słowo o słowie*, Warsaw 2009

Jan Grzenia, *Komunikacja językowa w Internecie,* Warsaw 2006

Marek Łaziński, *O panach i paniach*, Warsaw 2006

Słownictwo współczesnej polszczyzny w okresie przemian, ed. J. Mazur, 2000

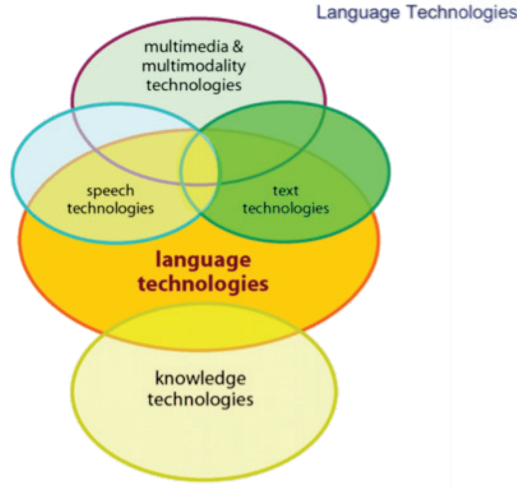
Zmiany w publicznych zwyczajach językowych, ed. J. Bralczyk, K. Mosiołek-Kłosińska, 2001

Computational Linguistics in Poland Web portal: http://clip.ipipan.waw.pl

Language Technology Support for Polish

Language Technologies

Language technologies are information technologies that are specialized for dealing with human language. Therefore these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under either speech or text technologies. Among those are technologies that link language to knowledge. The figure on the right illustrates the Language Technology landscape. In our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language and spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents.



Language Technology Application Architectures

Typical software applications for language processing consist of several components that mirror different aspects of language and of the task they implement. The figure on the right displays a highly simplified architecture that can be found in a text processing system. The first three modules deal with the structure and meaning of the text input:

- Pre-processing: cleaning up the data, removing formatting, detecting the input language and encoding, etc.
- Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- Semantic analysis: disambiguation (Which meaning of *apple* is the right one in a given context?), resolving anaphora and referring expressions like *she*, *the car*, etc.; representing the meaning of the sentence in a machine-readable way

Task-specific modules then perform many different operations such as automatic summarization of an input text, database lookups and many others. Below, we will illustrate **core application areas** and highlight their core modules. Again, the architectures of the applications are highly simplyfied and idealised, to illustrate the complexity of language technology applications in a generally understandable way.

After the introduction of the core application areas, we will give a short overview of the situation in LT research and education, concluding with an overview of (past) funding programs. In the end of this section, we will present an expert estimation on the situation regarding core LT tools and resources on a number of dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Polish.

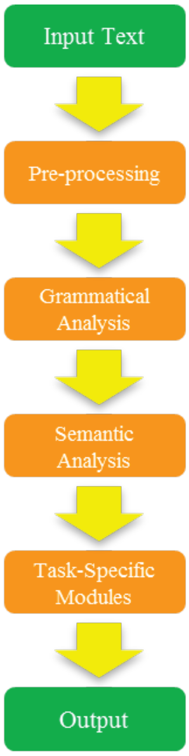


Figure 2: A Typical Text Processing Application Architecture

Core application areas

Language checking

Anyone using a word processing tool such as Microsoft Word has come across a **spell checking** component that indicates spelling mistakes and proposes corrections. 40 years after the first spelling correction program by Ralph Gorin, language checkers nowadays do not simply compare the list of extracted words against a dictionary of correctly spelled words, but have become increasingly sophisticated. In addition to language-dependent algorithms for handling morphology (e.g. plural formation), some are now capable of recognizing syntax–related errors, such as a missing verb or a verb that does not agree with its subject in person and number, e.g. in "She **write* a letter." However, for other common error types the above described methods are not sufficient. For example, if we take a look at the following first verse of a poem by Jerrold H. Zar (1992):

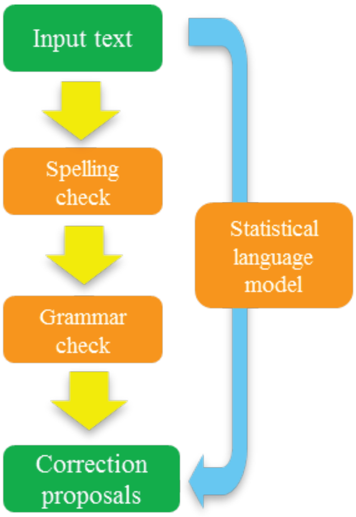


Figure 3: Language Checking (left: rule-based; right: statistical)

Eye have a spelling chequer,
It came with my Pea Sea.
It plane lee marks four my revue
Miss Steaks I can knot sea.

Most available spell checkers (including Microsoft Word) will find no errors in this poem because they mostly look at words in isolation. However, analysis of larger contexts is needed in many cases, e.g., for deciding if a word such as the adjective "polski" needs to be written in upper case, as in:

Ten tekst został przełożony na polski.
[This document was written in Polish.]
Czytał „Polskę Zbrojną".
[He read Polska Zbrojna.]

This either requires the formulation of language-specific grammar rules, i.e. a high degree of expertise and manual labour, or the use of a so-called statistical language model (alternatively, grammar rules might be induced using artificial intelligence methods). Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, "polska książka" is a much more probable word sequence than "Polska książka". A statistical language model can be automatically derived using a large amount of (correct) language data (i.e. a corpus). Up to now, these approaches have mostly been developed and evaluated on English language data. However, they do not necessarily transfer straightforwardly to Polish with its flexible word order and richer inflection. The rule-based methods have been used in the open-source proof-reading tool LanguageTool that incorporates over 1 thousand rules for Polish (the tool can be used in various word processing systems, such as LibreOffice).

The use of language checking is not limited to word processing tools, but it is also applied in **authoring support** systems. Accompanying the rising number of technical products, the amount of technical documentation has rapidly increased over the last decades. Fearing customer complaints about wrong usage and damage claims resulting from bad or badly understood instructions, companies have begun to focus increasingly on the quality of technical documentation, at the same time targeting the international market. Advances in natural language processing lead to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with certain rules and (corporate) terminology restrictions. As Polish is rarely a source language in such applications, no generic authoring system has been built especially for Polish.

Besides spell checkers and authoring support, language checking is also important in the field of computer-assisted language learning and is applied to automatically correct queries sent to web search engines, e.g. Google's 'Did you mean…' suggestions.

Web search

The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide[13]. Neither the search interface nor the presentation of the retrieved results has significantly changed since the first version. In the current version, Google offers a spelling correction for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix[14], which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a lot of data at hand and efficient techniques for **indexing** these data, a mainly statistically-based approach can lead to satisfactory results. (→information retrieval)

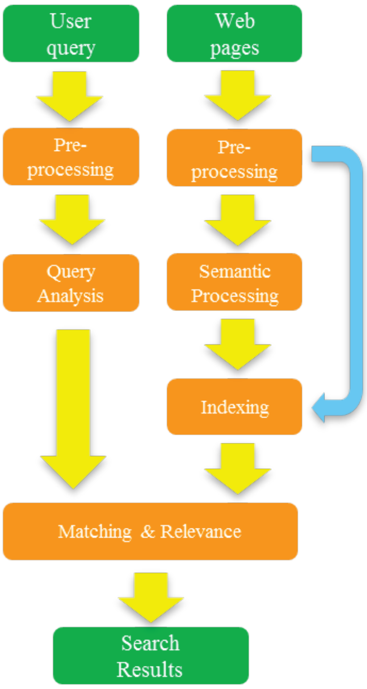


Figure 4: Web Search Architecture

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In research labs, experiments using machine-readable thesauri and ontological language resources like WordNet (or the equivalent Polish SłowoSieć) have shown improvements by allowing to find a page on the basis of synonyms of the search terms (e.g. “energia atomowa”, “energia jądrowa”, “energia nuklearna”, etc.) and even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology. If a search query consists of a question or another type of sentence rather than a list of keywords, retrieving relevant answers to this query requires an analysis of this sentence on a syntactic and semantic level as well as the availability of an index that allows for a fast retrieval of the relevant documents. For example, imagine a user inputs the query 'Give me a list of all companies that were taken over by other companies in the last five years'. For a satisfactory answer, one needs to apply a syntactic parser to analyse the grammatical structure of the sentence in order to determine that the user is looking for companies that have been taken over and not companies that took over others. One also needs to process the expression *last five years* to find out which years it refers to.

Finally, the processed query needs to be matched against a huge amount of unstructured data in order to find the piece or pieces of information the user is looking for. This involves the **retrieval** and **ranking** of relevant documents. In addition, generating a list of companies, we also need to extract the information that a particular string of words in a document refers to a company name. This kind of information is made available by so-called named-entity recognizers.

Even more demanding is the attempt to match a query to documents written in a different language. For **multilingual search**, we have to translate the query automatically to all possible source languages and transfer the retrieved information back to the target language. The increasing percentage of data available in non-textual formats drives the demand for services enabling **multi-media search**, i.e. information search on images, audio, and video data. For audio and video files, this involves a speech recognition module to convert speech content into text or a phonetic representation, to which user queries can be matched.

In Poland, SMEs like Carrot Search in Poznań successfully develop and apply search technologies that are able to provide more structured information than standard engines like Google by clustering the results in a language-sensitive way. Polish search engines include NetSprint and Szukacz. The latter contains a Polish thesaurus and stemmer, which enhances the search results.

Speech interaction

Speech interaction technology is the basis for the creation of interfaces that allow a user to interact with machines using spoken language rather than, e.g. a graphical display, a keyboard, and a mouse. Today, such voice user interfaces (VUIs) are usually employed for partially or fully automating service offerings provided by companies to their customers, employees, or partners via the telephone. Business domains that rely heavily on VUIs are banking, logistics, public transportation, and telecommunications. Other usages of speech interaction technology are interfaces to particular devices, e.g. in-car navigation systems, and the employment of spoken language as an alternative to the input/output modalities of graphical user interfaces, e.g. in smartphones.

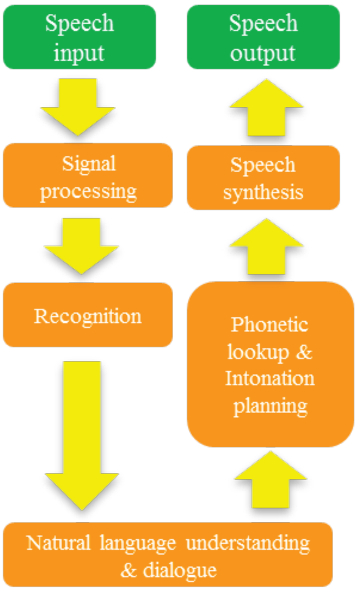

Speech input
Speech output
Signal processing
Speech synthesis
Recognition
Phonetic lookup & Intonation planning
Natural language understanding & dialogue

Figure 5: Simple Speech-based Dialogue Architecture

At its core, speech interaction comprises the following four different technologies:

- Automatic speech recognition (ASR) is responsible for determining which words were actually spoken given a sequence of sounds uttered by a user.
- Syntactic analysis and semantic interpretation deal with analysing the syntactic structure of a user's utterance and interpreting the latter according to the purpose of the respective system.
- Dialogue Management is required for determining, on the part of the system the user interacts with, which action shall be taken given the user's input and the functionality of the system.
- Speech Synthesis (Text-to-Speech, TTS) technology is employed for transforming the wording of that utterance into sounds that will be outputted to the user.

One of the major challenges is to have an ASR system recognise the words uttered by a user as precisely as possible. This requires either a restriction of the range of possible user utterances to a limited set of keywords, or the manual creation of language models that cover a large range of natural language user utterances. Whereas the former results in a rather rigid and inflexible usage of a VUI and possibly causes a poor user acceptance, the creation, tuning and maintenance of language models may increase the costs significantly. However, VUIs that employ language models and initially allow a user to flexibly express their intent – evoked, e.g., by a *How may I help you* greeting – show both a higher automation rate and a higher user acceptance and may therefore be considered as advantageous over a less flexible *directed dialogue* approach.

For the output part of a VUI, companies tend to use pre-recorded utterances of professional – ideally corporate – speakers a lot. For static utterances, in which the wording does not depend on the particular contexts of use or the personal data of the given user, this will result in a rich user experience. However, the more dynamic content an utterance needs to consider, the more user experience may suffer from a poor prosody resulting from concatenating single audio files. In contrast, today's TTS systems prove superior, though optimisable, regarding the prosodic naturalness of dynamic utterances.

Regarding the market for speech interaction technology, the last decade underwent a strong standardisation of the interfaces between the different technology components, as well as by standards for creating particular software artefacts for a given application. There also has been strong market consolidation within the last ten years, particularly in the field of ASR and TTS. Here, the national markets in the G20 countries – i.e. economically strong countries with a considerable population - are dominated by less than 5 players worldwide, with *Nuance* and *Loquendo* being the most prominent ones in Europe.

On the Polish TTS market, the most successful company is *Ivona* which offers products for other languages as well. However, for languages with a smaller number of speakers, commercially employable ASR and TTS products sometimes do not even exist. Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Today's key players in Poland are *PrimeSpeech* and *Skrybot*. Rather than exclusively relying on a product business based on software licenses, these companies have positioned themselves mostly as full-service providers that offer the creation of VUIs as a system integration service. Finally, within the domain of *speech* interaction, a genuine market for the linguistic core technologies for syntactic and semantic analysis does not exist yet.

As for the actual employment of VUIs, demand in Poland has strongly increased within the last 5 years. This tendency has been driven by end customers' increasing demand for customer self service and the considerable cost optimisation aspect of automated telephone services, as well as by a significantly increased acceptance of spoken language as a modality for man-machine interaction.

Looking beyond today's state of technology, there will be significant changes due to the spread of smartphones as a new platform for managing customer relationships – in addition to the telephone, Internet, and email channels. This tendency will also affect the employment of technology for speech interaction. On the one hand, demand for telephony-based VUIs will decrease, on the long run. On the other hand, the usage of spoken language as a user-friendly input modality for smartphones will gain significant importance. This tendency is supported by the observable improvement of speaker independent speech recognition accuracy for speech dictation services that are already offered as centralised services to smartphone users. Given this 'outsourcing' of the recognition task to the infrastructure of applications, the application-specific employment of linguistic core technologies will supposedly gain importance compared to the present situation.

Machine translation

The idea of using digital computers for translation of natural languages came up in 1946 by A. D. Booth and was followed by substantial funding for research in this area in the 1950s and beginning again in the 1980s. Nevertheless, machine translation (MT) still fails to fulfil the high expectations it gave rise to in its early years.

At its basic level, MT simply substitutes words in one natural language by words in another. This can be useful in subject domains with a very restricted, formulaic language, e.g., weather reports. However, for a good translation of less standardized texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty here lies in the fact that human language is ambiguous, which yields challenges on multiple levels, e.g., *word sense disambiguation* on the lexical level ('Jaguar' can mean a car or an animal) or the attachment of prepositional phrases on the syntactic level as in:

Policjant zauważył samochód w zaroślach.

[The policeman observed the car in the bush.]

Policjant zauważył samochód w okularach.

[The policeman observed the car through his glasses.]

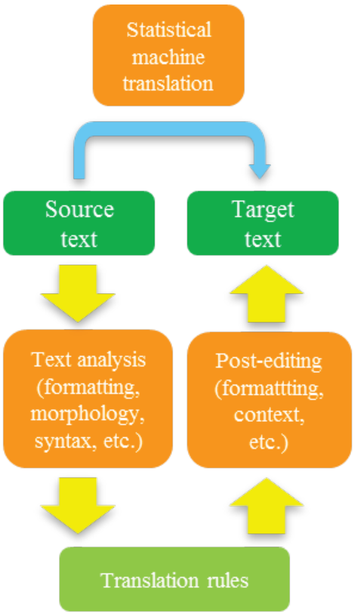

Statistical machine translation
Source text
Target text
Text analysis (formatting, morphology, syntax, etc.)
Post-editing (formatting, context, etc.)
Translation rules

Figure 6: Machine translation (top: statistical; bottom: rule-based)

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible in cases like the example above. But often **rule-based** (or knowledge-driven) systems analyse the input text and create an intermediary, symbolic representation, from which the text in the target language is generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by a skilled linguist.

Beginning in the late 1980s, as computational power increased and became less expensive, more interest was shown in **statistical models** for MT. The parameters of these statistical models are derived from the analysis of bilingual text corpora, such as the *Europarl* parallel *corpus*, which contains the proceedings of the *European Parliament* in 11 European languages. Given enough data, statistical machine translation works well enough to derive an approximate meaning of a foreign language text. However, unlike knowledge-driven systems, statistical (or data-driven) MT often generates ungrammatical output. On the other hand, besides the advantage that less human effort is required for grammar writing, data-driven MT can also cover particularities of the language that go missing in knowledge-driven systems, for example idiomatic expressions.

As the strengths and weaknesses of knowledge- and data-driven MT are complementary, researchers nowadays unanimously target **hybrid approaches** combining methodologies of both. This can be done in several ways. One is to use both knowledge- and data-driven systems and have a selection module decide on the best output for each sentence. However, for longer sentences, no result will be perfect. A better solution is to combine the best parts of each sentence from multiple outputs, which can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

For Polish, machine translation is challenging. The free word order poses problems for analysis, and extensive inflection is a challenge for generating words with proper gender and case markings.

The leading MT system for Polish is Translatica (Poleng) and it is widely available. Poleng works with the PWN Scientific Publishers and uses its extensive dictionaries, including the Oxford PWN English/Polish dictionary. Translatica is **rule-based** and supports Polish, English, German, and Russian. While there is significant research in this technology in national and international contexts, data-driven and hybrid systems have been less successful in business than in research so far.

However, generic **statistical** MT systems such as Google Translate and Bing support Polish to a considerable degree, especially in translation from and into English. Nevertheless, for other language pairs the performance is low and the results are far from understandable, sometimes even ridiculous. This is due to the scarcity of the parallel corpora that are used to train statistical MT.

Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly. Special systems for interactive translation support were developed e.g. at Poleng (TranslAide) and Studio Gambit (TIGER). There are also smaller SMEs offering Computer-Aided Translation (CAT) tools, such as Cafetran. A special MT system, Thetos, was built to translate Polish into sign language for the hearing impaired.

The quality of MT systems is still considered to have huge improvement potential. Challenges include the adaptability of the language resources to a given subject domain or user area and the integration into existing workflows with term bases and translation memories. In addition, most of the current systems are English-centred and support only a few languages combinations from and into Polish, which leads to frictions in the total translation workflow, and e.g. forces MT users to learn different lexicon coding tools for different systems.

Information management / Language Technology ‘behind the scenes’

Building Language Technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but provide significant service functionalities ‘under the hood’ of the system. Therefore they constitute important research issues that have become individual sub-disciplines of Computational Linguistics in academia.

Question answering has become an active area of research, for which annotated corpora have been built and scientific competitions have been started. The idea is to move from keyword-based search (to which the engine responds with a whole collection of potentially relevant documents) to the scenario of the user asking a concrete question and the system providing a single answer: ‘At what age did Neil Armstrong step on the moon?’ - ’38’. While this is obviously related to the aforementioned core area Web Search, question answering nowadays is primarily an umbrella term for research questions such as what *types* of questions should be distinguished and how should they be handled; how can a set of documents that potentially contain the answer be analysed and compared (do they give conflicting answers?); how can specific information - the answer - be reliably extracted from a document, without unduly ignoring the context; etc.

This is in turn related to the information extraction (IE) task, an area that was extremely popular and influential at the time of the 'statistical turn' in Computational Linguistics, in the early 1990s. IE aims at identifying specific pieces of information in specific classes of documents; this could be e.g. the detection of the key players in company takeovers as reported in newspaper stories. Another scenario that has been worked on is reports on terrorist incidents, where the problem is to map the text to a template specifying the perpetrator, the target, time and location of the incident, and the results of the incident. Domain-specific template-filling is the central characteristic of IE, which for this reason is another example of a 'behind the scenes' technology that constitutes a well-demarcated research area but for practical purposes then needs to be embedded into a suitable application environment.

Two 'borderline' areas, which sometimes play the role of standalone application and sometimes that of supportive, 'under the hood' component are text summarization and text generation. Summarization, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. It works largely on a statistical basis, by first identifying 'important"' words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarization equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarizers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesize *new* sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but is embedded into a larger software environment, such as the clinical information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

For Polish, the situation in all these research areas is much less developed than it is for English, where since the 1990s QA, IE, and summarization have been the subject of numerous open competitions, primarily those organized by DARPA/NIST in the United States. These have significantly improved the state of the art, but the focus has always been on English; some competitions have added multilingual tracks, but Polish was never prominent. Accordingly, there are hardly any annotated corpora or other resources for these tasks. Summarization systems, when using purely statistical methods, are often to a good extent language-independent, and thus some research prototypes are available. For text generation, reusable components have traditionally been limited to the surface realization modules (the "generation grammars"); again, most of the available software is designed for English. Prototype implementations of text generation were created during the development of MT system that translated Polish into sign language.

There are other fields in which linguistic technology is being applied. One of them is plagiarism detection, which uses language-independent technologies but may be enhanced with search for simple paraphrases of the text. The most popular Polish application in this field is the web-based system plagiat.pl, used in most higher education institutions to ensure originality of master's theses, as well as to detect document copyright infringement on the web.

LT Programs/Projects

One of the earliest significant projects in computational linguistics was the creation of the corpus of frequency dictionary of contemporary Polish by an interdisciplinary team of researchers from the University of Warsaw. The original purpose of the corpus was to create a general frequency dictionary of contemporary Polish. The work started in 1967. Partial results were published between 1972 and 1977, the completed dictionary in 1990. The corpus was later augmented in various respects, both by manual editing and automated procedures. Its design is comparable to the Brown corpus of English.

The early efforts included projects that aimed at the creation of a representative Polish morphological dictionary. One such project was POLEX (1993-1996) at Adam Mickiewicz University; another was Słownik Gramatyczny Języka Polskiego that resulted in the current state-of-the-art morphological analyzer for Polish, Morfeusz. In 2008, an important project plWordNet coordinated by Wrocław University of Technology (Institute of Applied Informatics), with the cooperation of Adam Mickiewicz University (POLNET project), was started in order to build the first Polish wordnet. The resulting wordnet is one of the biggest in the world (the coverage in some categories is larger than in Princeton Wordnet), and numerous innovative semi-automatic methods were used to discover meaning relations on the basis of linguistic corpora.

Another important corpus project was the IPI PAN corpus created in early 2000s at the Institute of Computer Science of the Polish Academy of Sciences (ICS PAS). It was the first comprehensible corpus to be available on the web for Polish.[15] At the same time, PWN scientific publishers developed their own corpus to be used for dictionary research, while at the University of Łódź, a corpus was built in the Pelcra project. In the next decade, a follow-up project, the National Corpus of Polish[16] was started by these three institutions and Institute of Polish Language (Cracow) and it already included some data from their existing resources. The goal of the project is to create the biggest Polish compiled from a pool of over 1 billion words with a manually annotated 1-million-word part (on several levels). These annotations will make it possible to prepare other linguistic resources from it. For example, a project was started to build the first Treebank for Polish using the grammatical annotations from the NKJP corpus.

Two projects in discourse processing, LUNA (ICS PACS) and POLINT-112-SMS (Adam Mickiewicz University) were started in the first decade of 2000s, to gather spoken language corpora and develop methods in discourse processing for Polish. The vision of LUNA was to improve automated telephone systems allowing easy human-machine interactions through spontaneous and unconstrained speech. POLINT-112-SMS is focused on information management in emergency situations. The input data for the system are human-generated text messages (SMS). They are processed to support decisions in a crisis management centre. One of the parts of the project is a dialog maintenance module.

Polish institutions are also involved in the ongoing CLARIN project and contribute to the efforts on the technological infrastructure for language resources and tools, and in FLaReNet, a European forum to facilitate interaction among language resources stakeholders. They are also active in META-NET project.

There are also at least 2 large ongoing projects financed by the EU under the Innovative Economy Programme (ATLAS and NEKST), and numerous other research projects in language technology, including the ones in the Framework Programme.

More financial means are necessary to support projects aiming at developing more sophisticated LT, language corpora and other language resources.

LT Research and Education

Poland has a number of excellent centres active in the field of language technology and computational linguistics. Currently, at least 12 Polish universities and research centres are active in the field. Many of them offer courses in the field of language technology.[17]

Apart from the universities, major research projects are carried out by the language technology group of the Institute of the Computer Sciences of the Polish Academy of Sciences (ICS PAS).

Polish associations active in the field of language technology are-Polskie Towarzystwo Informatyczne and Polskie Towarzystwo Fonetyczne.

LT as a field of research faces the following problems:

- Since researchers are part of different communities they meet in several separate conferences and have different meetings and boards. Hence, there is no single conference at which one can meet all stakeholders.
- Computational linguistics is still seen as an ‘exotic’ topic, which has not acquired a fixed place in the faculty system yet, and hence is located in different faculties, e.g. the computer science faculties or in the humanities.
- Research topics dealt with are overlapping only partially.
- Challenges (e.g. CLEF/textual entailment): good results but no influential community (rather success of individuals)

Availability of tools and resources

The following table provides an overview of the current situation of language technology support for Polish. The rating of existing tools and resources is based on educated estimations by several leading experts using the following criteria (each ranging from 0 to 6).

1 **Quantity**: Does a tool/resource exist for the language at hand? The more tools/resources exist, the higher the rating.

- 0: no tools/resources whatsoever
- 6: many tools/resources, large variety

2 **Availability**: Are tools/resources accessible, i.e.,are they Open Source, freely usable on any platform or only available for a high price or under very restricted conditions?

- 0: practically all tools/resources are only available for a high price
- 6: a large amount of tools/resources is freely, openly available under sensible Open Source or Creative Commons licenses that allow re-use and re-purposing

3 **Quality**: How well are the respective performance criteria of tools and quality indicators of resources met by the best available tools, applications or resources? Are these tools/resources current and also actively maintained?

- 0: toy resource/tool
- 6: high-quality tool, human-quality annotations in a resource

4 **Coverage**: To which degree do the best tools meet the respective coverage criteria (styles, genres, text sorts, linguistic phenomena, types of input/output, number languages supported by an MT system etc.)? To which degree are resources representative of the targeted language or sublanguages?

- 0: special-purpose resource or tool, specific case, very small coverage, only to be used for very specific, non-general use cases
- 6: very broad coverage resource, very robust tool, widely applicable, many languages supported

5 **Maturity**: Can the tool/resource be considered mature, stable, ready for the market? Can the best available tools/resources be used out-of-the-box or do they have to be adapted? Is the performance of such a technology adequate and ready for production use or is it only a prototype that cannot be used for production systems? An indicator may be whether resources/tools are accepted by the community and successfully used in LT systems.

- 0: preliminary prototype, toy system, proof-of-concept, example resource exercise
- 6: immediately integratable/applicable component

6 **Sustainability**: How well can the tool/resource be maintained/integrated into current IT systems? Does the tool/resource fulfil a certain level of sustainability concerning documentation/manuals, explanation of use cases, front-ends, GUIs etc.? Does it use/employ standard/best-practice programming environments (such as Java EE)? Do industry/research standards/quasi-standards exist and if so, is the tool/resource compliant (data formats etc.)?

- 0: completely proprietary, ad hoc data formats and APIs
- 6: full standard-compliance, fully documented

7 **Adaptability**: How well can the best tools or resources be adapted/extended to new tasks/domains/genres/text types/use cases etc.?

- 0: practically impossible to adapt a tool/resource to another task, impossible even with large amounts of resources or person months at hand
- 6: very high level of adaptability; adaptation also very easy and efficiently possible

Table of Tools and Resources

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology (Tools, Technologies, Applications)							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	4	5	5	5	5	4	4
Parsing (shallow or deep syntactic analysis)	4	4	4	4	3	4	2
Sentence Semantics (WSD, argument structure, semantic roles)	1	2	4	1	1	2	4
Text Semantics (coreference resolution, context, pragmatics, inference)	1	0	3	1	0	1	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	1	0	1	1	0	1	0
Information Retrieval(text indexing, multimedia IR, crosslingual IR)	2	5	2	2	4	4	4
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	3	3	2	2	1	3	4
Language Generation (sentence generation, report generation, text generation)	1	1	1	1	1	1	2
Summarization, Question Answering, advanced Information Access Technologies	1	1	2	2	1	1	1
Machine Translation	3	4	3	3	3	4	3
Speech Recognition	1	2	3	4	3	2	4
Speech Synthesis	4	3	6	5	4	4	3
Dialogue Management (dialogue capabilities and user modelling)	1	1	1	1	1	1	1

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Resources (Resources, Data, Knowledge Bases)							
Reference Corpora	3	2	4	4	5	5	3
Syntax-Corpora (treebanks, dependency banks)	2	1	4	4	1	4	4
Semantics-Corpora	1	0	4	2	2	0	4
Discourse-Corpora	1	1	2	1	1	1	1
Parallel Corpora, Translation Memories	3	1	4	4	5	5	5
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	1	0	3	3	2	2	2
Multimedia and multimodal data (text data combined with audio/video)	1	1	1	1	1	0	0
Language Models	1	3	1	1	1	1	1
Lexicons, Terminologies	3	2	4	4	4	4	2
Grammars	3	2	4	4	3	2	2
Thesauri, WordNets	3	4	4	4	3	4	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	2	1	4	2	1	1	2

Note: in the above table, parsing includes shallow and deep syntactic analysis. It should be stressed, however, that for Polish, practical parsing is now possible only on the shallow level, while deep parsing is only a single prototype (Świgra parser based on Świdziński's grammar).

Conclusions

The table can be summarized in the form of a number of key messages, which highlight crucial issues for the further development of automatic language processing of Polish on the basis of the present situation:

- For Polish, discourse corpora or advanced discourse processing are not widely available. Multimodal corpora are in preparation.
- Many of the resources lack standardization, i.e. even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.
- Semantics is more difficult to process than syntax; text semantics is more difficult to process than word and sentence semantics.

- The more semantics a tool takes into account, the more difficult it is to find the right data; more efforts for supporting deep processing are needed.
- Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.
- Speech processing, specially speech synthesis, is currently more mature than NLP for written text.
- Research was successful in designing particular high quality software, but it is nearly impossible to come up with sustainable and standardized solutions given the current funding situations.
- Polish lacks large, balanced and more easily available parallel corpora, including large parallel corpora for related languages such as Czech or Polish.
- For many purposes, bilingual and multilingual dictionaries that include not only translations but also valency information seem indispensable. These need to be built, as standard dictionaries usually omit this kind of annotation.
- Large and widely available ontological resources for Polish are needed for many applications. Currently available ontologies are relatively small, based on OpenCyc or on Polish OpenThesaurus. A Polish version of DBPedia is in preparation.

About META-NET

META-NET is a Network of Excellence funded by the European Commission. The network currently consists of 47 members from 31 European countries. META-NET fosters the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe.

Language Users
Society
Research Communities
Industries

The Multilingual Europe Technology Alliance (META)

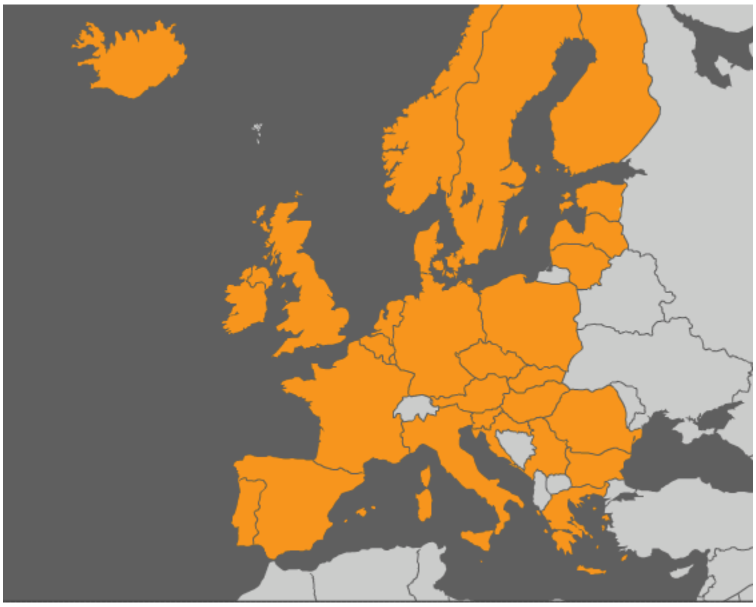


Figure 1: Countries Represented in META-NET

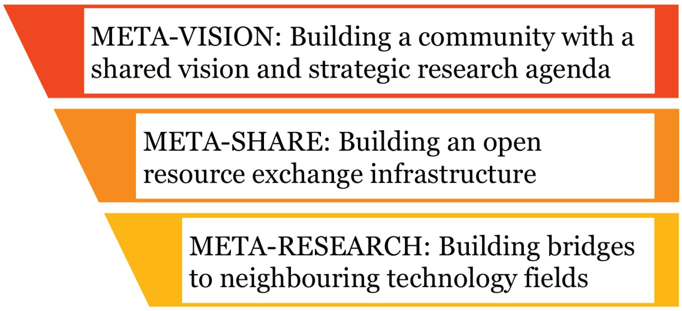
META-NET cooperates with other initiatives like the Common Language Resources and Technology Infrastructure (CLARIN), which is helping establish digital humanities research in Europe. META-NET fosters the technological foundations for the establishment and maintenance of a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- provides equal access to information and knowledge in any language;
- offers advanced and affordable networked information technology to European citizens.

META-NET stimulates and promotes multilingual technologies for all European languages. The technologies enable automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. The network wants to improve current approaches, so better communication and cooperation across languages can take place. Europeans have an equal right to information and knowledge regardless of language.

Lines of Action

META-NET launched on 1 February 2010 with the goal of advancing research in language technology (LT). The network supports a Europe that unites as a single, digital market and information space. META-NET has conducted several activities that further its goals. META-VISION, META-SHARE and META-RESEARCH are the network's three lines of action.



META-VISION: Building a community with a shared vision and strategic research agenda

META-SHARE: Building an open resource exchange infrastructure

META-RESEARCH: Building bridges to neighbouring technology fields

Figure 2: Three Lines of Action in META-NET

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. In the first year of META-NET, presentations at the FLaReNet Forum (Spain), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to develop its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET communicated the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items. META-SHARE targets existing language data, tools and systems as well as new and emerging products that are required for building and evaluating new technologies, products and services. The reuse, combination, repurposing and re-engineering of language data and tools plays a crucial role. META-SHARE will eventually become a critical part of the LT marketplace for developers, localisation experts, researchers, translators and language professionals from small, mid-sized and large enterprises. META-SHARE addresses the full development cycle of LT—from research to innovative products and services. A key aspect of this activity is establishing META-SHARE as an important and valuable part of a European and global infrastructure for the LT community.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the Semantic Web community. META-RESEARCH focuses on collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. This activity has already clearly identified aspects of MT where semantics can impact current best practices. In addition, the activity has created recommendations on how to approach the problem of integrating semantic information in MT. META-RESEARCH is also finalising a new language resource for MT, the Annotated Hybrid Sample MT Corpus, which provides data for English-German, English-Spanish and English-Czech language pairs. META-RESEARCH has also developed software that collects multilingual corpora that are hidden on the web.

Member Organisations

The following table lists the organisations and their representatives that participate in META-NET.

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	University of Zagreb	Marko Tadić
Cyprus	University of Cyprus	Jack Burston
Czech Republic	Charles University in Prague	Jan Hajic
Denmark	University of Copenhagen	Bolette Sandford Pedersen and Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi and Krister Linden
France	CNRS/LIMSI	Joseph Mariani
	Evaluations and Language Resources Distribution Agency	Khalid Choukri
Germany	DFKI	Hans Uszkoreit and Georg Rehm
	RWTH Aachen University	Hermann Ney
	Saarland University	Manfred Pinkal
Greece	Institute for Language and Speech Processing, "Athena" R.C.	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi

Country	Organisation	Participant(s)
	Budapest University of Technology and Economics	Géza Németh and Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Italy	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	Institute of Mathematics and Computer Science, University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Utrecht University	Jan Odijk
	University of Groningen	Gertjan van Noord
Norway	University of Bergen	Koenraad De Smedt
Poland	Polish Academy of Sciences	Adam Przepiórkowski and Maciej Ogrodniczuk
	University of Lodz	Barbara Lewandowska-Tomaszczyk and Piotr Pęzik
Portugal	University of Lisbon	Antonio Branco
	Institute for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	Alexandru Ioan Cuza University	Dan Cristea
Serbia	University of Belgrade	Dusko Vitas, Cvetana Krstev and Ivan Obradovic
	Institute Mihailo Pupin	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	Pompeu Fabra University	Núria Bel

Country	Organisation	Participant(s)
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiadou
	University of Edinburgh	Steve Renals

References

[1] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[2] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[3] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[4] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[5] http://cdt.europa.eu/EN/whoweare/Pages/OurEUlanguages.aspx

[6] http://www.efnil.org/documents/language-legislation-version-2007/poland/poland

[7] http://browse.oecdbookshop.org/oecd/pdfs/free/9810081e.pdf

[8] http://www.rp.pl/artykul/645517.html

[9] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[10] http://www.dns.pl/zonestats.html

[11] http://www.ebrandservices.com/welcome-to-e-brand-services,130.html

[12] http://ling.pl/

[13] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[14] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

[15] http://korpus.pl

[16] http://www.nkjp.pl

[17] http://clip.ipipan.waw.pl/Centers