META-NET White Paper Series

Languages in the European Information Society

– Dutch –

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

Prof. Dr. Jan Odijk, Utrecht University

Contributors

Drs. Alice Dijkstra, NOW
Prof.dr. Jean-Pierre Martens, Ghent University
Dr. Jacomine Nortier, Utrecht University
Dr. Peter Spyns, Dutch Language Union
Drs. Remco van Veenendaal, Dutch HLT Agency

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 Dutch language demonstrates that a lively language technology industry and research environment exists in the Netherlands and Flanders. Although a number of technologies and resources for Standard Dutch exist, there are fewer technologies and resources for the Dutch language than for the English language. The technologies and resources also have a poorer quality.

According to the assessment detailed in this report, immediate action must occur before any breakthroughs for the Dutch 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 currently witnessing a digital revolution that is comparable to Gutenberg's invention of the printing press.

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?

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.

Dutch in the European Information Society

General Facts

With about 23 million native speakers, Dutch is the 8th most widely spoken native language in the EU. It is the commonly used language in the Netherlands and the Flemish part (called Flanders) of Belgium and one of the official languages in Surinam, Aruba, Curacao and Sint-Maarten, where it is used by parts of the population. It is also spoken in the EU in France and Germany, and outside the EU in Brazil, Canada, Indonesia (Java and Bali), South Africa, and the United States. The official Dutch name for the language is *Nederlands*, though Dutch as spoken in Flanders is usually called *Vlaams* ('Flemish').

This White Paper focuses on the situation of the Dutch language and LT for it in the Netherlands and Flanders, which together we will designate with the term 'the Low Countries'.

In the Netherlands, Dutch is the common spoken and written language and the native language of the vast majority of the population. The Netherlands has one officially recognized minority language, Frisian, spoken in the province of Friesland (Frisia). There are several immigrant languages. No reliable figures on the number of speakers of immigrant languages are known. However, the *Centraal Bureau voor de Statistiek* (Statistics Netherlands)[5] does provide figures for immigrants by ethnicity (=/= nationality). For ethnicities from outside the Netherlands some 1.5 million are from Western origin, and for non-western origin the figures are: Morocco (Rif Berber, estimated at 75%, and (Moroccan) Arabic, estimated at 25%) 350k persons, Netherlands Antilles and Aruba (Papiamentu) 138k persons, Surinam (Dutch, Sranan, Guyanese Creole English, Hindustani, Javanese) 342k persons, Turkey (Turkish) 383k persons, and other non-western (various languages) 644k persons.

In Belgium, Dutch is, by law, the language of Flanders, and one of the two languages (next to French) of the Brussels region. Belgium also has a French-speaking region and a German-speaking region.

Dutch has a variety of dialects, including (in the Netherlands) Achterhoeks, Drents, Gronings, Limburgs, Sallands, Stellingwerfs, Twents, Veluws and Zeeuws, and in Flanders West-Vlaams, Antwerps, Oost-Vlaams, Brabants and Limburgs. The orthography is standardized but there were some changes in the standard recently (1996 and 2006). The standard is obligatory in education and governmental publications. Some of the recently proposed changes have led to different interpretations of the standard by different publishers, causing small differences in spelling (e.g. the *Groene Boekje*[6]*: actievoeren v. Van Dale: actie voeren*), and some spelling changes were not accepted by all publishers[7], who spell certain words differently (esp. with regard to the so-called *tussen-n* in compounds), in accordance with the so-called *Witte Boekje*.[8] Dutch orthography can be quite complicated for certain words and constructions, so complicated that every year the so-called *Groot Dictee*[9] is organized by the Netherlands and Flanders and broadcast on national TV. The *Groot Dictee* is so difficult that anyone scoring less than 30 errors in about 8 sentences can be considered an excellent speller! In general, all Dutch dialects in the Netherlands share the same core grammar, though some dialects exhibit

differences in some syntactic constructions. There are several lexical differences between dialects, and especially between Dutch as spoken in the Netherlands and Dutch as spoken in Flanders, e.g. the word “ajuin” is only used in Flanders instead of the standard Dutch “ui” (‘onion’). There are also several words that are the same in Flanders and in the Netherlands but have a different meaning, e.g. ‘middag’ (lit. ‘midday’) in the Netherlands means the period of the day from 14:00-17:00 hrs, while in Flanders it means the period of the day from 12:00-14:00. Flemish also uses many words originating from French, e.g. terms for car engine parts, while Dutch in the Netherlands uses more English or English-inspired words in this domain. This also sometimes has consequence for pronunciation, e.g. the words *flat* and *tram* are in use both in the Netherlands and in Flanders, they are borrowed from English but in Flanders the borrowing went via the French language, so that in Flanders these words are pronounced as fl[A]t and tr[A]m while in the Netherlands they are pronounced as fl[E]t and tr[E]m

Particularities of the Dutch Language

The Dutch language exhibits some specific characteristics, which contribute to the richness of the language by allowing the speakers to express ideas in a large variety of ways. One such particularity is that it is quite common to put non-subjects sentence-initially (much more common than in English). For example, consider the English sentence

The woman was going to the store every day.

In English, there are very limited possibilities to use a different word order in this sentence, but in the Dutch equivalent almost any phrase can be the initial phrase in the sentence:

De vrouw ging elke dag naar de winkel

Elke dag ging de vrouw naar de winkel

Naar de winkel ging de vrouw elke dag

Word order in Dutch is thus much freer than in English (but not as free as in German).

Also, the Dutch language is quite productive in creating new compounds, though the use and productivity of compounding is not as extreme as in German. Nevertheless, newly formed compounds occur frequently and are difficult to process for NLP technology.

An other characteristics of Dutch that makes processing difficult is formed by separable verb prefixes that can occur far from the verb in nested constructions like:

*Hij **stelde** zich na mij een drankje aangeboden te hebben en wij in gesprek geraakt waren aan ons **voor***

*(He **introduced** himself after he offered me a drink and we started a conversation.)*

The meaning of a verb containing such a separable prefix like “voor”, “in” or “uit” can very often not be derived from the meaning of the base verb and the meaning of the prefix. For example, the verb “stellen” (put, place), is contained in “voorstellen” (imagine/introduce/etc.), “instellen” (set up/regulate/etc.), “uitstellen” (postpone.) and many other verbs.

A further peculiarity complicating automatic processing of Dutch is the phenomenon of so-called R-pronouns such as *er, waar, daar*. These pronouns are often at a distance from the preposition that they belong to

*Hij keek **daar** gisteren **naar***

*(he was looking **at that** yesterday)*

where *daar* and *naar* are separated from each other by the adverb *gisteren* 'yesterday'. Furthermore, a single occurrence of the R-pronoun *er* can serve multiple functions at once, e.g. in

Dachten er twee over na?

(Were two of them thinking about it?)

where *er* belongs both to the preposition *over* 'about' and the quantifier *twee* 'two'.

Recent developments

From the 1950s on, American TV series and movies began to conquer the Dutch market. Foreign films and series are generally broadcast in the original language and subtitled. The strong presence of the American way of life in the media influenced the Dutch culture and language. Due to the continuing triumph of English music since the 1960s (e.g. Elvis Presley, the Beatles), generations of young people grew up naturally surrounded by English. The English language rose to become the "cool/hip" language and has kept this status until today.

The lasting popularity is expressed by the fact that nowadays loan words often originate from the English language. According to an estimate by [Van der Sijs 2005], 30% of the Dutch vocabulary are loan words, and many of these are English loan words. In most cases these words fill some gap, i.e., they enrich the Dutch language rather than threaten it, though some are considered *anglicisms*, i.e. barbarisms from the English language for which proper Dutch equivalents exist which should preferably be used.

Borrowings from English are dominating in business, science, certain technical domains and on the internet. A strong tendency to overuse English loan words can also be detected in product advertisements.

These developments demonstrates the importance of raising awareness for a development that entails the risk of excluding large parts of the population from taking part in information society, namely those who are not familiar with English. They were one of the reasons to set up the Dutch-Flemish language and speech technology programme STEVIN[10], which aimed to consolidate the position of the Dutch language in the modern information society.

Language cultivation in the Low Countries

The Dutch language is represented by various publicly funded societies and language bodies. One example is the Nederlandse Taalunie (Dutch Language Union)[11], in which the Netherlands, Flanders and Surinam cooperate on the Dutch language. It deals with the Dutch language itself, the Dutch language in digital applications, education in and of the Dutch language, literature, the promotion of reading skills, and the position of the Dutch language in Europe and the world.

Private initiatives include *het Genootschap Onze Taal*[12] ('Society of Our Language'), and het *Algemeen Nederlands Verbond*[13] (General Dutch Union).

There are several institutes dedicated to the study of the Dutch language and culture, e.g. *het Instituut voor Nederlandse Lexicologie (INL, 'Institute for Dutch Lexicology')*[14], the *Meertens Institute* (that studies the Dutch language and its dialects and Dutch culture)[15], and the *Huygens ING Institute* (for the study of Dutch literature and history)[16]. The latter two are institutes of the *Koninklijke Nederlandse Academie voor Wetenschappen* (KNAW, Royal Netherlands Academy of Arts and Sciences).[17] Furthermore, the *TST-Centrale (Dutch HLT-Agency)*[18], which is part of INL, stores, maintains and distributes HLT-resources for the Dutch language.

Unlike some other countries, the Netherlands does not maintain a language academy, but Belgium does have the *Koninklijke Academie voor Nederlandse Taal- en Letterkunde*[19] (Royal Academy of Dutch Literature and Linguistics).

Measures to protect the status of the Dutch language are rarely taken. One exception is the 'language laws' set up in Belgium, with its complicated and sensitive language situation, in part to protect Dutch against French. In the area of language technology, the funding of the STEVIN programme to consolidate the position of the Dutch language in the modern information and communication society is rare and only short-term exception, and the set-up of the TST-Centrale (Dutch HLT Agency) a good (but very small) step towards a more long term approach.

The Dutch language is relatively small, and its native speakers are generally well-educated to speak other languages (esp. English), which puts the Dutch language in disadvantageous situation compared to e.g. languages like French, which have a large speaker basis and which is strongly promoted by the global community of French-speaking peoples within the so-called Francophonie. These factors may encourage an attitude of tolerance and openness towards cultural diversity, but can also pose a threat to Dutch language cultivation.

Language in Education

The Ministry of OCW (Education, Culture and Sciences) organizes and monitors education in general, including the education of the Dutch language. Language skills are the key qualification needed in education as well as for personal and professional communication. Dutch language teaching makes up about one third of the school lessons of 9-to-11-year-old students, comparable the native language lessons in France and Greece and higher than the 20% reported for Germany. It is therefore not surprising that, on a European level, the PISA 2009 study revealed that Dutch students performed significantly above OECD average with respect to reading literacy.[20]

The education of Dutch 'extra muros' is also systematically monitored via studies performed by or under the supervision of the Dutch Language Union.[21] The Dutch Language Union focus involves not only research but also concrete policy and practical guidelines for addressing problems in areas such as spelling, reading skills, language competence of teachers, language and/or educational retardation, education in literature, and others.

Continuous attention to Dutch language teaching in schools is essential for providing students with the language skills required for an active participation in society. Language technology can make an important contribution here by offering so-called computer-assisted language learning (CALL) systems, which allow students to experience language in a playful way, for example by linking special vocabulary in electronic text to comprehensible definitions or to audio or video files supplying additional information, e.g., the pronunciation of a word.

International aspects

The Dutch language has produced authors of international standing, and many authors reach an international audience via translations of their works.[22] Nevertheless, its influence is small in comparison to big languages such as English, German and French. In philosophy, the Netherlands has made significant contributions (e.g. Spinoza, and more recently (in the area of the foundations of mathematics) L.E.J. Brouwer and E.W. Beth). The Low Countries have a flourishing scientific community and a high international prestige. Eighteen scientists from the Netherlands and 5 from Belgium (of which 2 from Flanders) have won Nobel prizes in physics, chemistry, economy, literature and medicine.

The Dutch language has never played an important role in international scientific publications. Though many publications on Dutch law, literature and history are written in Dutch, most scientific publications are in English. In many conferences, workshops and lectures at Dutch universities the working language is English. This is also true in the business world. In many large and internationally active companies, English has become the lingua franca, both in written (emails and documents) and oral communication (e.g. talks).

Even though Dutch is taught by 700 teachers at 190 universities and by 6000 teachers to 400,000 students at hundreds of non-university institutes, the status of Dutch as a foreign language has always been marginal in comparison to big languages such as English. Pragmatic reasons for learning Dutch (e.g. better chances on the job market) are of little importance, so most students must be driven by pure interest in the Dutch language.

Within the European Union, Dutch is an official language, but Dutch is hardly used in European Union business. Only the official legislation, some documents for Dutch-speaking members of the European parliament, and documents aimed at the general public are published also in Dutch, turning Dutch into a somewhat marginal language at the EU level, and endangering the interest of the Dutch speaking communities.

Language technology can address this challenge from a different perspective by offering services like machine translation or cross-lingual information retrieval to foreign language text and thus help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

Dutch on the Internet

In June 2010, 88.6% of the Dutch[23] were internet users and 72.7% of the Flemish[24] had internet. Among young people, the proportion of users is even higher. There is an active Dutch-speaking web community, e.g. reflected by the Dutch Wikipedia, the ninth largest Wikipedia in the world.[25] A recent study showed that 90% of the

European internet users prefer reading a website in their native language over reading a website in a non-native language[26], and only a small minority would accept a web page in English if there is no alternative in their own language. Furthermore, active use of the internet drops to 35% when it has to be done in a non-native language. This witnesses to the importance of the native language on the internet.

With about 1.24 million Internet domains[27], the Netherlands's top-level country domain .nl is the 11th country extension. Though not bad for a small country and growing, the amount of Dutch language data available on the web is of course minor compared to the English language data and language data from several other bigger languages such as German and French.

For language technology, the growing importance of the internet is important in two ways. 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 involving 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 in the second part of this paper. It involves sophisticated language technology, differing for each language. For Dutch, this comprises matching words with variants with changed spellings as well as words with diacritics such as accents and tremas with words without these diacritics. But internet users and providers of web content can also profit from language technology in less obvious ways, for example if it is used to automatically translate web contents from one language into another. Considering the high costs associated with manually translating these contents, it may be surprising how little usable language technology is built in compared to the anticipated need.

However, it becomes less surprising if we consider the complexity of (the Dutch) 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 Dutch.

Selected Further Reading

Languages of the Netherlands, Ethnologue
http://www.ethnologue.com/show_country.asp?name=nl

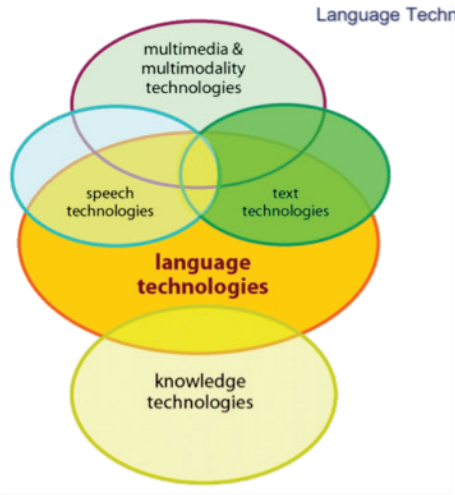
Dutch, Ethnologue.
http://www.ethnologue.com/show_language.asp?code=nld

E-ANS http://www.let.ru.nl/ans/e-ans/index.html

Language Technology Support for Dutch

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:

deal with the structure and meaning of the text input:

- Pre-processing: cleaning up the data, removing formatting, detecting the input language, replacing "e" by "ë" for Dutch, 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 the given context?), resolving the reference of pronouns such as "she", and expressions such as "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 look-ups and many others. Below, we will illustrate **core application areas** and highlight certain of the modules of the different architectures in each section. Again, the architectures are highly simplified and idealised, serving for illustrating the complexity of language technology applications in a generally understandable way.

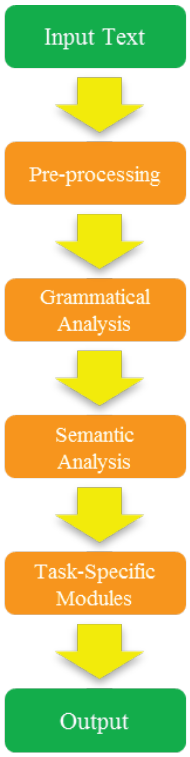


Figure 2: A Typical Text Processing Application Architecture

After the introduction of the core application areas, we will shortly give an 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 in a number of

dimensions such as availability, maturity, or quality. This table gives a good overview on the situation of LT for Dutch.

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, most available spell checkers (including Microsoft Word) will find no errors in the following first verse of a poem by Jerrold H. Zar (1992):

Eye have a spelling chequer,

It came with my Pea Sea.

It plane lee marks four my revue

Miss Steaks I can knot sea.

For handling this type of errors, analysis of the context is needed in many cases, e.g., for deciding whether a verb has to be written with *dt* or *d at the end* in Dutch, as in:

Hij heeft het dier verwond.

[He has injured the animal]

Hij verwondt het dier.

[He injures the animal.]

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. Such models calculate the probability of a particular word occurring in a specific environment (i.e., the preceding and following words). For example, *hij verwondt* is a much more probable word sequence than *hij verwond*. 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 Dutch with its more flexible word order, verb particle combinations, compounds, and crossing dependencies.

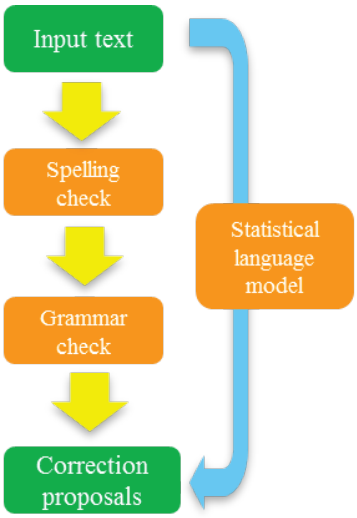


Figure 3: Language Checking (left: rule-based; right: statistical)

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 led 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.

Proofing tools for Dutch that were incorporated in Microsoft products were developed in the past by Lernout & Hauspie, independently later by Polderland, and they are currently maintained and further developed by Knowledge Concepts. Other companies active in this area are TALŌ BV and Carp technologies.

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

Search on the web, in intranets, or in digital libraries is probably the most widely used and yet underdeveloped Language Technology today. The search engine Google, which started in 1998, is nowadays used for about 80% of all search queries world-wide[28]. The verb *googelen* even has an entry in the Dutch *Van Dale* dictionary. 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[29], 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.

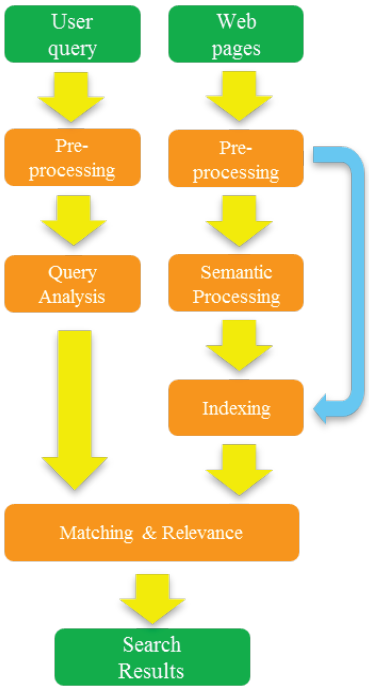


Figure 4: Web Search Architecture

However, for a more sophisticated request for information, integrating deeper linguistic knowledge is essential. In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet (or the equivalent Dutch EuroWordNet), have shown improvements by allowing to find a page on the basis of synonyms of the search terms, e.g. *kernenergie* and *nucleaire energie* (atomic energy, nuclear energy) or 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, syntactic parsing needs to be applied to analyse the grammatical structure of the sentence and determine that the user is looking for companies that have been taken over and not companies that took over others. Also, the expression *last five years* needs to be processed in order 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 is commonly referred to as information retrieval and involves the search for 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 cross-lingual information retrieval, we have to automatically translate the query 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 multimedia information retrieval, 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 the Netherlands, several companies are active in these domains, including AskNow Solutions, Carp Technologies, GridLine, Irion Technologies, Knowledge Concepts, MediaLab Solutions, RightNow! (formerly Q-Go), TextKernel, and others. In Belgium Natlanco, InterSystems (formerly i.Know), ICMS, Aktor Technologies, Mentoring Systems and CrossMinder are active in these areas.

The focus of development for these companies is in providing add-ons and advanced search engines for special-interest portals by exploiting topic-relevant semantics. Due to the still high demands in processing power, such search engines are only economically usable on relatively small text corpora. Processing time easily exceeds that of a common statistical search engine as, e.g., provided by Google by a magnitude of thousands. These search engines also have high demand in topic-specific domain modelling, making it not feasible to use these mechanisms on web scale.

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 smart phones.

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 output 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 is difficult

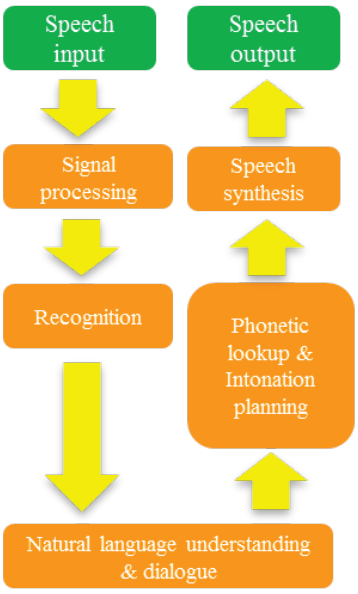


Figure 5: Simple Speech-based Dialogue Architecture

because speech does not contain spaces between words (as in written language), and because the speech signal is highly variable in character (accent differences, male voices differ from female voices, background noises, etc.). 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 his/her 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 the user experience may suffer from a poor prosody resulting from concatenating single audio files. In contrast, today's TTS systems prove superior regarding the prosodic naturalness of dynamic utterances, even though improvements are still possible.

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. Nuance has a big development centre in Flanders.

On the Dutch TTS market, there are additional smaller companies like Acapela, based in Wallonia, SVOX, headquartered in Switzerland, and Fluency, based in Amsterdam. There are many companies that are active in using TTS and ASR technology in applications and services. These include Advance Voice Technology, Dialogs Unlimited, DutchEar, G2 Speech, Logica, OrcaVoice, Quentris, Sensotec NV, Telecats, TomTom and Voice Data Bridge. Several companies and foundations focus on applications for user groups with specific demands such as physically handicapped people, dyslectic people, and elderly. These include Axendo, Cochlear Benelux, Dedicon, JABBLA, Kamelego, Lexima, rdgKompagne and VoiceCore

Regarding dialogue management technology and know-how, some relevant companies are Carp technologies, Irion, RightNow! (formerly Q-Go) and RE-Phrase for text-based applications, and Dialogs Unlimited, DutchEar, Telecats, and Voice Data Bridge for speech-based applications. 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 has increased within the last 5 years. This tendency has been driven by end customers' increasing demand for customer self-service and the con-

siderable 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 smart phones 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 smart phones 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 (e.g. *graven* can mean 'counts', 'graves' or 'to dig') or the interpretation of relative pronouns (as subject or as object) on the syntactic level as in:

De man die de vrouw zag

[The man who saw the woman] or

[The man who the woman saw]

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 MT works well enough to derive an approximate meaning of a

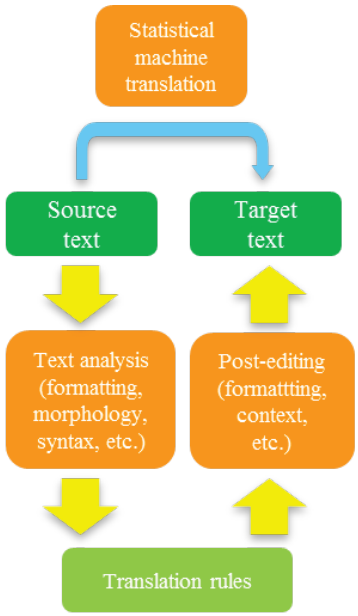


Figure 6: Machine translation (top: statistical; bottom: rule-based)

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 Dutch, MT is particularly challenging. The possibility of creating arbitrary new words by compounding makes dictionary analysis and dictionary coverage difficult; rather free word order, split verb constructions and R-pronouns pose problems for analysis.

Leading commercial MT systems like Systran, Globalink, LOGOS, METAL (and its spin-offs, LANT, GMS and Lucy Software), LMT developed by IBM (forming the basis for Linguatec and Lingenio) never covered the Dutch language, probably because it was not interesting to do so from a commercial point of view. Only some research systems for Dutch were developed, partially in companies (Philips: Rosetta, BSO: Distributed Translation) and partially in academia (Utrecht University: Eurotra). Translation systems for Dutch were only produced when funded. For example, METAL produced a Dutch-French MT system for the ministries of Agriculture and Internal Affairs, and after the Dutch Language Union issued a call for the development of MT systems translating between Dutch on the one hand and English and French on the other in 1999,[30] funded by public money, Systran developed such systems in the context of the NL-Translex project.

All systems mentioned above were knowledge-based. With the rise of statistical MT, Dutch has become a language quite generally covered. It is included in the 52 languages Google Translate offers and in the 24 languages SDL Language Weaver offers.

Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly. Most MT companies stress that they can rapidly adapt their standard systems to company-specific dictionaries, terminology and translation memories, thereby increasing MT quality significantly.

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 few languages directly from and into Dutch, which leads to frictions in the total translation workflow, and e.g. forces MT users to learn different lexicon coding tools for different systems.

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?), and how can specific information - the answer - be reliably extracted from a document, without unduly ignoring the context.

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, is used in virtually every search engine to provide a snippets of a found document, 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 embedded into a larger software environment, such as into the clinical

information system where patient data is collected, stored and processed, and report generation is just one of many functionalities.

For Dutch, the situation in all these research areas is much less developed than it is for English, where question answering, information extraction, and summarization have since the 1990s 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 Dutch was never prominent, though some challenges are organized from Flanders.[31] Nevertheless, work on question answering was promoted by the IMIX programme that focused on Interactive Multimodal Information eXtraction applied to Dutch resources.[32] In this programme, question answering systems, with speech input and output, supporting follow-up questions were developed for the general domain and one specific for the medical domain, as well as systems to generate textual output in combination with other modalities, and dialogue managers to connect these systems. Furthermore, the company RightNow (formerly Q-GO) from the Netherlands has been very successful in the area of textual question answer systems operating via chats or e-mail. Eindhoven University (IPO) has worked on a language and speech generation system, that has later been acquired by Polderland (and probably now resides with Knowledge Concepts), but it appears hardly to have been used outside its original purpose.[33] Tilburg University has worked on multi-document summarization (integrating different messages on the same topic) in the STEVIN DAESO project.[34] Nevertheless, there are hardly any annotated corpora or other resources for these tasks.

LT Research and Education

In academia there are a number of excellent centres in the area of human language technology, e.g. KU Leuven, Ghent university, Radboud University Nijmegen and University of Twente for speech technology, Tilburg and Antwerp university for machine learning techniques in NLP, Utrecht University, and Leuven for NLP and machine translation, Groningen and Amsterdam for parsing, Amsterdam for sentiment mining and parsing, etc. It is, however, very difficult to attract students for the NLP field. Possible causes for this may be the relative low visibility of NLP in the university curricula and the fact that many NLP research groups are in the humanities (students there do not easily take a technical view on language, as is required for NLP).

The academic players in the Netherlands and Flanders do not necessarily focus on the Dutch language: in research the focus is typically on English in order to be able to make sensible comparisons with researchers abroad. Nevertheless, several researchers are active in the area of Computer Aided Language Learning (CALL), where language and speech technology is used to increase language skills of first and second language learners. Relevant organizations include RU Nijmegen, University of Antwerp Linguapolis and KU-LAK.

LT Industry and Programs

The LT field in the Netherlands and Belgium consist of many organisations, both industry (some 65) and knowledge centres (44).[35] The sector is reasonably well organized, with an active professional

organization NOTaS[36] in the Netherlands consisting of 15 industrial and academic partners, the Flemish research community cooperating in CLIF[37], and intense cooperation in the last decade between players from the Netherlands and Flanders, and from industry and academia in the joint Netherlands-Flanders LT programmes CGN (Spoken Dutch Corpus)[38] and especially STEVIN. The SMEs in Flanders, however, are acting individually, and have not organized themselves in a sector, which makes them relatively invisible.

Most industrial players are very small SMEs and have to struggle every day to survive, or they are small departments in a company that has a different focus for its core business activities. Nevertheless, some SMEs are quite successful and have been able to build up a stable business. Most SMEs in the area of speech technology are system integrators, application developers, or service providers. The actual development of technology, at least in speech technology, has been concentrated in a very few number of players (e.g. Nuance).

One problem for marketing LT is that LT is not clearly visible because it is hidden as an integrated part of a more encompassing product or service, even though it is a component of products and services used by many users (e.g. search on the internet, texting on mobile phones, etc.).

Even though there are many players in the Netherlands and Flanders, this does not imply that their focus is also on the Dutch language. For industry, the Dutch language is commercially less interesting than other languages, and the necessary investments can often not be justified by the small Dutch-language market.

Language Technology Programs

Activities for the Dutch language have to be promoted and supported explicitly. Fortunately, this has been done in several programmes and projects over the last one and a half decade. Thus a Dutch language spoken train information system was developed as a carrier for research in speech analysis and generation, in language analysis and generation, and in dialogue management in the OVIS programme in the late nineties. The NL-Translex project was already mentioned above. Flanders had a short term programme on LT in the mid nineties. The IMIX programme, mentioned above, carried out research using systems for the Dutch language. In the IOP MMI (Innovation Research Programme on Man Machine Interaction) and CATCH[39] programs language and speech technology have been used as tools for man machine interfaces and disclosing cultural heritage. Most prominent in their focus on the Dutch language are the joint Netherlands-Flanders CGN and STEVIN programmes. These have yielded significant progress in the availability of basic resources (data and tools) for the Dutch language, some initial research and several end user applications. Though some of the results achieved in these projects can be exploited in industry and in academia (e.g. in the CLARIN-NL project[40]) the prospects for optimally exploiting these results in actual research and in industry further are grim, since it appears not to have the focus of attention of the government in the Netherlands, and research has been reorganized so that it has become more difficult to get funding for discipline-specific programmes. The situation is probably a bit more positive in Flanders, though. Furthermore, some prerequisites for exploiting the potential are in place,

such as visibility and accessibility of the resources produced in earlier programmes via the TST-Centrale (Dutch HLT Agency).

The programmes mentioned also contributed significantly to bringing together the speech and language technology communities, which until recently were heterogeneous communities and operated quite separated from each other. These disciplines are distributed over computer or engineering science faculties (speech technology in Flanders, and in Twente; some language technology) and the humanities faculties (most though not all language technology) and generally meet in several separate conferences. The only exception may be the LREC conference[41], which however has a specific focus on language resources and evaluation.

It is generally expected that the role of LT is going to be boosted enormously by the increasing growth of content that is ubiquitously available via an increasing amount of small mobile devices with large computational power (smart phones, iPad, etc.) and continuous access to the internet. Such devices have a relatively small screen, and no or primitive keyboards, which makes the use of speech increasingly more natural and necessary, and the amount of information they must search, summarize, translate or otherwise process requires an enormous boost in LT technology.

Status of Tools and Resources for Dutch

The following table provides an overview of the current situation of language technology support for Dutch. The rating of existing technologies 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

Status of Tools and Resources for Dutch

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

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

Conclusions

The situation of Dutch concerning language technology support gives rise to cautious optimism. Supported by larger research programs in the past, there exists a language technology industry and research scene in the Low Countries, which consists mostly of SMEs but is already partially organized.

For standard Dutch, a number of technologies and resources exist, but far less than for English. As has been shown by several past studies on specific areas of language technology such as EuroMatrixPlus, Dutch plays in Europe's second league together with German and French and few other languages. Still, this information has to be taken with care as even for English, language technology support today is by far not in a state that is needed for offering the support a true multilingual knowledge society needs.

In this Whitepaper Series, the first effort has been made to assess the overall situation of many European languages with respect to language technology support in a way that allows for high level comparison and identification of gaps and needs.

For Dutch, key results regarding technologies and resources include the following:

- Speech processing currently seems to be more mature than processing of written text (though more complex applications still pose serious challenges to speech technology).
- Advanced information access technologies are in their infancies (Information Extraction, Question Answering, Advanced Discourse Processing, Summarization, etc.).
- The more linguistic and semantic knowledge a tool takes into account, the more gaps exist (see, e.g., information retrieval vs. text semantics); more efforts for supporting deep linguistic processing are needed.
- Research was successful in designing particular high quality software, but many of the resources lack standardization and especially interoperability; concerted programs and initiatives are needed to make data and tools truly interoperable.
- For Dutch, many resources created with public money in the recent LT programmes are either open source or stored, maintained and distributed by the HLT Agency and easily and cheaply accessible. (cf. the high scores for Availability for Parsing, Sentence semantics, Text semantics, Reference corpora, Syntax corpora and Semantic corpora)
- Annotated corpora with semantic structures are available but minimal in size and depth of annotation. Annotated corpora with discourse structures are lacking almost completely, as are Ontological resources for World Knowledge.
- Parallel corpora for machine translation are available but in quantities that are too small for proper development of MT systems. MT, and especially statistical MT, needs huge amounts of (parallel) data to perform reasonably.
- Multimedia data is a huge gap.

From this, it is clear that more efforts need to be directed into the creation of resources for Dutch and into research, innovation, and development. The need for large amounts data and the high complexity of language technology systems make it also mandatory to develop new infrastructures for sharing and cooperation.

Bibliography

[Cucchiarini et al 2000] Catia Cucchiarini, Johan Van Hoorde and Elizabeth D'Halleweyn (2000), 'NL-Translex: Machine Translation for Dutch', Proceedings of LREC 2000. http://www.mt-archive.info/LREC-2000-Cucchiarini.pdf

[Ethnologue 2009] Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com

[Dutch Wikipedia: Nederlands] Dutch Wikipedia, entry for *Nederlands* http://nl.wikipedia.org/wiki/Nederlands

[Dutch Wikipedia: Nedersaksisch] Dutch Wikipedia, entry for Nedersaksisch *http://nl.wikipedia.org/wiki/Nedersaksisch*

[Dutch Wikipedia: Fries] *Dutch Wikipedia, entry for Fries (spoken in the Netherlands*
http://nl.wikipedia.org/wiki/Westerlauwers_Fries

[Internet World Stats] Internet World Stats, Copyright © 2010, Miniwatts Marketing Group. http://www.internetworldstats.com

[Joscelyne & Lockwood 2003] Andrew Joscelyne and Rose Lockwood (2003) "Benchmarking HLT progress in Europe, the EUROMAP study", Copenhagen 2003.
http://www.csc.fi/yhteistyo/tulokset/2003/euromap_report

[Soria & Mariani 2011] Claudia Soria & Joseph Mariani (2011): "Report on Existing Projects and Initiatives", META-NET study, 2011.

[Taalunieversum Facts and Trivia] *Taalunieversum web page on the Dutch language: Facts and Trivia*
http://taalunieversum.org/taal/feiten_en_weetjes/

[Theune 2003] Theune, M. (2003), 'Natural Language Generation for dialogue: system survey', accessible via CiteSeer
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.3398&rep=rep1&type=pdf

[Van der Sijs 2005] Nicoline van der Sijs, *Groot leenwoordenboek*, Utrecht/Antwerpen 2005

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.

The Multilingual Europe Technology Alliance (META)

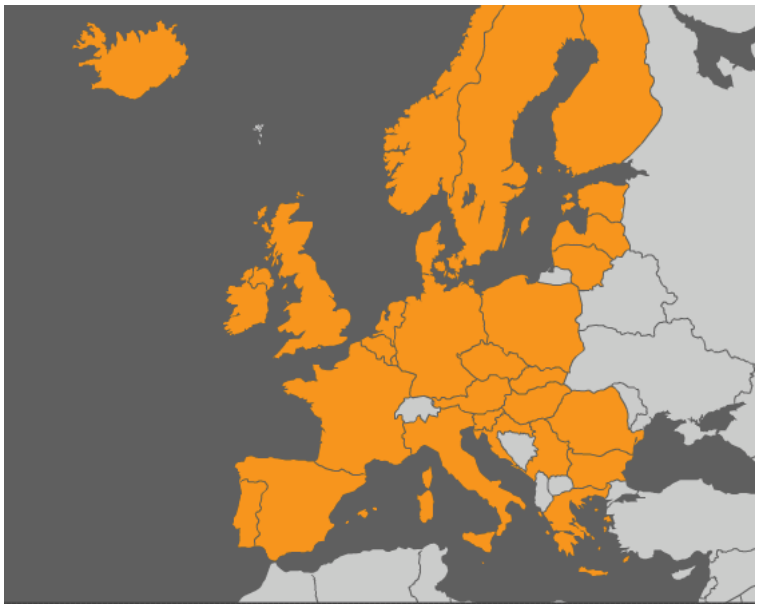


Figure 1: Countries Represented in META-NET

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.

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 collect-

ing 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://www.cbs.nl/nl-NL/menu/home/default.htm?Languageswitch=on

[6] http://nl.wikipedia.org/wiki/Groene_Boekje

[7] http://www.onzetaal.nl/dossier/spelling/wittespellers.php

[8] http://www.onzetaal.nl/advies/wittespelling.php

[9] http://grootdictee.nps.nl/

[10] http://taalunieversum.org/taal/technologie/stevin/

[11] http://taalunieversum.org/taalunie/

[12] http://www.onzetaal.nl/ot/index.php

[13] http://www.algemeennederlandsverbond.org/

[14] http://www.inl.nl/

[15] http://www.meertens.knaw.nl/cms/

[16] http://www.huygensinstituut.knaw.nl/

[17] http://www.knaw.nl/

[18] http://www.inl.nl/tst-centrale

[19] http://www.kantl.be/

[20] http://www.oecd.org/dataoecd/54/12/46643496.pdf

[21] http://taalunieversum.org/onderwijs/algemeen/

[22] http://www.nlpvf.nl/vertalingendb/search1.php

[23] http://www.internetworldstats.com/stats4.htm

[24] VRIND 2010, p. 188. http://www4.vlaanderen.be/dar/svr/Pages/2010-10-28-vrind2010.aspx

[25] http://meta.wikimedia.org/wiki/List_of_Wikipedias#All_Wikipedias_ordered_by_number_of_articles

[26] http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf

[27] http://www.webhosting.info/domains/country_stats/NL

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

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

[30] [Cucchiarini et al. 2000]

[31] http://webs.hogent.be/~elef464/lt3_SemEval.html

[32] http://www.nwo.nl/imix

[33] LGM, see [Theune 2003]

[34] http://daeso.uvt.nl/

[35] See http://taalunieversum.org/taal/technologie/organisaties/ for a pretty complete overview.

[36] http://www.notas.nl/

[37] http://clif.esat.kuleuven.be/

[38] http://lands.let.kun.nl/cgn/

[39] http://www.nwo.nl/catch

[40] http://www.clarin.nl

[41] http://www.lrec-conf.org/