

**METANET4U** 

**D2.3.en**  
**Language Report for**  
**English**

Version 1.1

2011-06-29



# METANET4U

[www.metanet4u.eu](http://www.metanet4u.eu)

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.

This central objective is articulated in terms of the following main goals:

**Assessment:** to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.

**Collection:** to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.

**Distribution:** to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.

**Dissemination:** to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.



## Deliverable D2.3.en: Language Report for English

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission



and by the participating institutions:



Faculty of Sciences, University of Lisbon



Instituto Superior Técnico



University of Manchester



University *Alexandru Ioan Cuza*



Research Institute for Artificial Intelligence,  
Romanian Academy



University of Malta



Technical University of Catalonia



Universitat Pompeu Fabra

## Deliverable D2.3.en: Language Report for English

### Revision History

Version	Date	Author	Organisation	Description
1.1	29-06-2011	Sophia Ananiadou, Paul Thompson, John McNaught	UNIMAN	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.



**METANET4U**

**D2.3.en**  
**Language Report for**  
**English**

Document METANET4U-2011-D2.3.en  
EC CIP project #270893

Deliverable  
Number: D2.3.en  
Completion: Final  
Status: Submitted  
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: University of Manchester  
Authors: Sophia Ananiadou, Paul Thompson, John McNaught  
Reviewers: Mike Rosner, Jan Joachimsen

© all rights reserved by FCUL on behalf of METANET4U



## Table of Contents

<b>Executive Summary .....</b>	<b>3</b>
<b>A Risk for Our Languages and a Challenge for Language Technology .....</b>	<b>5</b>
Language Borders Hinder the European Information Society.....	5
Our Languages at Risk.....	6
Language Technology is a Key Enabling Technology .....	6
Opportunities for Language Technology .....	7
Challenges Facing Language Technology.....	8
Language Acquisition .....	8
<b>English in the European Information Society.....</b>	<b>10</b>
General Facts .....	10
Particularities of the English Language .....	10
Recent developments.....	11
Language cultivation in the UK.....	12
Language in Education.....	13
International aspects .....	14
English on the Internet.....	14
Selected Further Reading.....	15
<b>Language Technology Support for English.....</b>	<b>16</b>
Language Technologies .....	16
Language Technology Application Architectures .....	16
Core application areas.....	17
<i>Language checking</i> .....	17
Web search .....	18
Speech interaction.....	19
<i>Machine translation</i> .....	21
Language Technology ‘behind the scenes’ .....	23
Language Technology in Education .....	24
Language Technology Programs .....	24
Status of Tools and Resources for English .....	26
Conclusions .....	27
<b>About META-NET.....</b>	<b>30</b>
Lines of Action .....	30
Member Organisations .....	32
<b>References .....</b>	<b>35</b>





## 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 English language demonstrates that a lively language technology industry and research environment exists in English-speaking countries, both in Europe and worldwide.

Although English is the language on which most language technology research has been carried out, the assessment detailed in this report reveals that there is still a large number of issues that must be addressed in order for English language technology to reach its full potential.

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



## A Risk for Our Languages and a Challenge for Language Technology

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

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

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

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

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

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

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

### Language Borders Hinder the European Information Society

We cannot precisely know what the future information society will look like. When it comes to discussing a common European energy strategy or foreign policy, we might want to listen to European foreign ministers speak in their native language. We might want a

platform where people, who speak many different languages and who have varying language proficiency, can discuss a particular subject while technology automatically gathers their opinions and generates brief summaries. We also might want to speak with a health insurance help desk that is located in a foreign country.

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

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.<sup>i</sup> 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?”

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

*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.<sup>ii</sup> 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.<sup>iii</sup>

*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.<sup>iv</sup> 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:

- ❑ 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

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

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 Europe can also help us communicate with our global partners and their multilingual communities. Language technologies support a wealth of international economic opportunities.

### 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.

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.

Learning a second language usually requires much more effort when a child is not immersed in a language community of native

*Multilingualism is the rule, not an exception.*

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

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

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.

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 have 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.

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

# English in the European Information Society

## General Facts

Around the world, there are around 375 million native speakers of English. As such, it is estimated to be the third largest language, coming behind only Mandarin Chinese and Spanish. English is a (co)-official language in 53 countries worldwide.

Within Europe, English is the most commonly used language in the United Kingdom. It is not an official language, since the UK does not have a formal constitution. However, it can be considered the *de facto* language, given that it is the official language of the British government, and is spoken by around 94 % of the 62 million inhabitants of the UK<sup>v</sup>. It is also the most widely spoken language in the Republic of Ireland (population approx 4.5 million), where English is the second official language (Irish being the first). English is additionally the official language of Gibraltar (a British overseas territory) and a co-official language in Jersey, Guernsey and the Isle of Man (British Crown Dependencies), as well as in Malta. Outside of Europe, the countries with the greatest number of native English speakers are the United States of America (215 million speakers), Canada (17.5 million speakers) and Australia (15.5 million speakers).

In addition to English, the UK has further recognised regional languages, according to the European Charter for Regional or Minority Languages (ECRML), i.e., Welsh, Scottish Gaelic, Cornish, Irish, Scots, and its regional variant Ulster Scots. Since February 2011, the Welsh language (which is spoken by approximately 20% of the population of Wales) has shared official status with English in Wales<sup>vi</sup>. The large number of British Asians (approx 2.3 million or 4% of the population, according to the 2001 census) give rise to other languages being spoken in the UK, most notably Punjabi and Bengali.

Due to global spread of English, a large number of dialects have developed. Major dialects such as American English and Australian English can be split into a number of sub-dialects. In recent times, differences in grammar between the dialects have become relatively minor, with major variations being mainly limited to pronunciation and, to some extent, vocabulary, e.g., *bairn* (child) in northern England and Scotland.

In addition to dialects, there are also a number of English-based pidgins and creole languages. Pidgins are simplified languages that develop as a means of communication between two or more groups that do not have a language in common. An example is Nigerian pidgin, which is used as a *lingua franca* in Nigeria, where 521 languages have been identified. A creole language is a pidgin that has become nativised (i.e. learnt as a native language), such as Jamaican Patois.

## Particularities of the English Language

Compared to most European languages, English has minimal inflection, with a lack of grammatical gender or adjectival agreement. Grammatical case marking has also largely been abandoned, with personal pronouns being a notable exception, where nominative case (I, we, etc.), accusative/dative case (me, us, etc.) and genitive case (my, our, etc.) are still distinguished.



A particular feature of the English language is its spelling system, which is notoriously difficult to master for non-native speakers. Whilst in many languages, there is a consistent set of rules that map spoken sounds to written forms, this is not the case in English. Nearly every sound can be spelt in more than one way, and conversely, most letters can be pronounced in multiple ways. Consequently, English has been described as “the world’s worst spelled language”.<sup>vii</sup>

Consider the /u:/ sound, which in English can be spelt (among other ways) as “oo” as in *boot*, “u” as in *truth*, “ui” as in *fruit*, “o” as in *to*, “oe” as in *shoe*, “ou” as in *group*, “ough” as in *through* and “ew” as in *flew*. Having multiple written ways to represent a single sound is not in itself an unusual feature of written languages. For example, the same sound can be written in French as “ou”, “ous”, “out” or “oux”. However, what is more unusual about English is the fact that most of the written forms have alternative pronunciations as well, e.g. *rub*, *buil*d, *go*, *toe*, *out*, *rough*, *sew*. One of the most notorious amongst the groups of letters listed is *ough*, which can be pronounced in up to ten different ways.

These special features of English are the result of a number of factors, including the complex history of the UK, which has been heavily influenced by previous invasions and occupations by Scandinavians and Normans. Also, English spelling does not reflect the significant changes in the pronunciation of the language that have occurred since the late fifteenth century. In contrast to many other languages, and despite numerous efforts, most efforts to reform English spelling have met with little success.

A further defining feature of English is the large number of phrasal verbs, which are combinations of verb and preposition and/or adverb. The meaning of phrasal verbs is often not easily predictable from their constituent parts, which make them an obstacle for learners of English. By means of an example, the verb *get* can occur in a number of phrasal verb constructions, such as *get by* (cope or survive), *get over* (recover from) and *get along* (be on good terms)

## Recent developments

Events in the more recent history of the UK have had a significant influence on the vocabulary of English. These events include the industrial revolution, which necessitated the coining of new words for things and ideas that had not previously existed, and the British Empire. At its height, the empire covered one quarter of the earth’s surface, and a large number of foreign words from the different countries entered the language. The increased spread of public education increased literacy, and, combined with the spread of public libraries in the 19<sup>th</sup> century, books (and therefore a standard language) were exposed to a much greater number of people. The migration of large numbers of people from many different countries to the United States of America also affected the development of American English.

The two world wars of the 20<sup>th</sup> century caused people from different backgrounds to be thrown together, and the increased social mobility that followed contributed to many regional differences in the language being lost, at least in the UK. With introduction of radio broadcasting, and later of film and television, people were further exposed to unfamiliar accents and vocabulary, which also influenced the development of the language. Today, American Eng-

lish has a particularly strong influence on the development of British English, due to the USA's dominance of cinema, television, popular music, trade and technology (including the Internet).

The online edition of the Oxford English Dictionary is updated four times per year, with the March 2011 release including 175 new words, many of which indicate the rapidly changing nature of our society<sup>viii</sup>. These words include initialisms such as *OMG* (Oh my god) and *LOL* (Laughing out loud), which reflect the increasing influence of electronic communications (e.g., email, text messaging, social networks, blogs, etc.) on everyday lives. An increasing thirst for travel and cuisines of the world has caused loan words such as *banh mi* (Vietnamese sandwich) to be listed.

Within Europe, English can today be considered the most commonly used language, with 51% of EU citizens speaking it either as a mother tongue or a foreign language, according to EUROBAROMETER survey<sup>ix</sup>. Considering non-native speakers of English in the EU, 38% state that they have sufficient English skills to hold a conversation. Of the 29 countries polled, English is the most widely known language apart from the mother tongue in 19 of these countries, with particularly high percentages of speakers in Sweden (89%), Malta (88%) and the Netherlands (87%).

### Language cultivation in the UK

There are a number of associations, both nationally and internationally, which aim to promote the English language. These include the English Association<sup>x</sup>, which was founded in 1906, with the aim of furthering knowledge, understanding and enjoyment of the English language and its literature, and to foster good practice in its teaching and learning at all levels. The Council for College and University English<sup>xi</sup> and the National Association for the Teaching of English<sup>xii</sup> promote standards of excellence in the teaching of English at different levels, from early years through to university studies. The European Society for the Study of English<sup>xiii</sup> promotes the study and understanding of English languages, literature, and cultures of English-speaking people within Europe.

The Queen's English Society<sup>xiv</sup> (QES) is a charity founded in 1972, which aims to protect the English language from perceived declining standards. Its objectives include the education of the public in the correct and elegant usage of English, whilst discouraging the intrusion of anything detrimental to clarity or euphony. Such intrusions include the introduction of "foreign" words and, in recent years, new technologies such as internet chat and text messaging. As such, the aims of the QES appear to be in conflict with those of the Oxford English Dictionary, which aims to describe recent changes in the language, rather than taking a prescriptive view of what is correct.

The aims of the QES are not so different from those of the language academies that exist in other European countries (e.g., L'Académie Française in France, the Real Academia Española in Spain and the Accademia della Crusca in Italy). These academies determine standards of acceptable grammar and vocabulary, as well as adapting to linguistic change by adding new words and updating the meanings of existing ones. Indeed, in 2010, it was attempted to form an Academy of English using a similar model to the academies listed above. However, such a prescriptive approach generated a large amount of bad press concerning objections to the suppression of linguistic diversity and evolution. Consequently, the project was abandoned after a few months.

## Language in Education

From the early 1960s until 1988, there was little or no compulsory English grammar teaching in schools. The Education Reform act of 1988, and with it the introduction of the National Curriculum, has resulted in greater structure in the teaching of English in the UK, including the re-introduction of grammar as a required element. From ages 5 – 16, during which the study of English is a compulsory subject (except in Wales), the teaching requirements are divided into the key areas of listening, speaking, reading and writing<sup>xv</sup>. The study of language structure, as well both standard English and variations (including dialects), and culture, are an integral part of each of the key areas, and are developed throughout the learning process. Between 2003 and 2010, the study of a foreign language was only compulsory between the ages of 11 – 14, causing a 30% drop in the number of students opting to study a foreign language beyond 14. However, from 2010, foreign language learning was planned to begin at age 10.

From the age of 16, education in the UK is optional. A 2006 survey of subjects studied by 16-18 year olds in England found that English literature was the third most popular subject (after General Studies and Mathematics)<sup>xvi</sup>, studied by approximately 19.5% of students. In contrast, only 7% per cent of students opt to study English language, making it the 14th most popular subject. This still puts it above the two most popular foreign languages, i.e. French at 22<sup>nd</sup> position (5% of students) and German at 29<sup>th</sup> position (2% of students). At degree level in UK universities, English ranked as the 6<sup>th</sup> most popular subjects in 2010, with a small increase in applications (8.6%) compared to 2009.

PISA studies<sup>xvii</sup> measure reading literary skills amongst teenagers in different countries. According to the results, UK students are failing to improve at the same rate as students in some other countries. Although the overall scores of UK teenagers have not altered significantly between 2000 and 2009, their performance compared to other participating countries has dropped from 7<sup>th</sup> to 25<sup>th</sup> position. According to the amount spent per student on teaching, the UK ranks 8<sup>th</sup> among the 65 countries taking part. The overall literacy score for the UK is not statistically significant from the average score of all participant countries, and as such has comparable rates of teenage literacy to countries such as France, Germany and Sweden and Poland. In the 2009 study, around 18% of UK students did not achieve the basic reading level.

In PISA studies, a major factor influencing reading performance variability between schools was found to be the socio-economic background of the students. The UK has quite a large percentage of immigrant students, with around 200 different native languages being represented at British schools<sup>xviii</sup>. However, there is generally a small gap between the performance of natives and immigrants. Although immigrants who do not speak English at home have considerably reduced skills, children whose native language is not English receive linguistic support to enable them to attain the minimum level of understanding and expression to follow their studies.

Within Europe, English is the most studied foreign language within schools, with a study carried out by Eurydice<sup>xix</sup> revealing that 90% of all European pupils learn English at some stage of their education. It is the mandatory first foreign language in 13 countries of Europe.

## International aspects

Driven by both British imperialism and the ascension of the USA as a global superpower since the Second World War, English has been increasingly developing as the *lingua franca* of global communication. It is the dominant or even the required language of communications, science, information technology, business, aviation, entertainment, radio and diplomacy, and a working knowledge of English has become a requirement in a number of fields, occupations and professions such as medicine and computing. As a consequence of this, over a billion people now speak English, at least to a basic level. Within the European Union, English is one of the three working languages of the European Commission (together with French and German). It is also one of the six official languages of the United Nations.

In science, the dominant nature of English can be viewed in two ways. On the one hand, its use as a common language in scientific publishing allows for ease of information storage and retrieval, and for knowledge advancement. On the other hand, English can be seen as something of a *Tyrannosaurus rex* - “a powerful carnivore gobbling up the other denizens of the academic linguistic grazing grounds”<sup>xx</sup>. Scientists face a great deal of pressure to publish in visible (usually international) journals, most of which are now in the English language, leading to a self-perpetuating cycle in which English is becoming increasingly important.

The global spread of English is creating further negative impacts, e.g., the reduction of native linguistic diversity in many parts of the world. Its influence continues to play an important role in language attrition.

## English on the Internet

In 2010, 30.1 million adults in the UK (approximately 60%) used the Internet almost daily, which is almost double the estimate of 2006<sup>xxi</sup>. The same report found that 19.1 million UK households (73%) had an Internet connection. It was found that Internet use is linked to various socio-economic and demographic indicators. For example, 60% of users aged 65 or over had never accessed the Internet, compared to 1% of those ages 16 to 24. Educational background also has an impact on Internet use. Some 97% of degree-educated adults had used the Internet, compared to 45% of people without formal qualifications.

In 2010, there were an estimated 536 million users of the English language Internet, constituting 27.3% of all Internet users<sup>xxii</sup>. This makes the English Internet the most used in the world, with only the Chinese Internet coming anywhere close, with 445 million users. The third most popular language is Spanish, with about 153 million users.

With 9.1 million registrations in February 2011, the UK’s top-level country domain, *.uk*, is the fifth most popular extension in the world. It is also the second most used country-specific extension, beaten only by the Germany’s *.de* extension<sup>xxiii</sup>.

For language technology (LT), 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 for which language technology is applicable.

The most commonly used web application is 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 English, this may consist of matching spelling variations (e.g. British/American variations such as *colour/color*), or using context to distinguish whether the word *fly* refers to a noun (insect) or verb.

It is an expressed political aim in the UK and other European countries to ensure equal opportunities for everyone. In particular, the *Disability Discrimination Act*, which came into force in 1995, together with the more recent *Equality Act* of 2010, have made it a legal requirement for companies and organisations to ensure that their services and information are accessible to all. This requirement applies directly to websites and Internet services. User-friendly language technology tools offer the principal solution to satisfy this legal regulation, for example by offering speech synthesis for the blind.

Internet users and providers of web content can also profit from language technology in less obvious ways, for example, in the automatic translation of 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 English language, which has been partially highlighted above, 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 English.

### Selected Further Reading

David Crystal: *The English Language: A Guided Tour of the Language*, Penguin, 2002

David Crystal: *Evolving English: One Language, Many Voices*, The British Library Publishing Division, 2010.

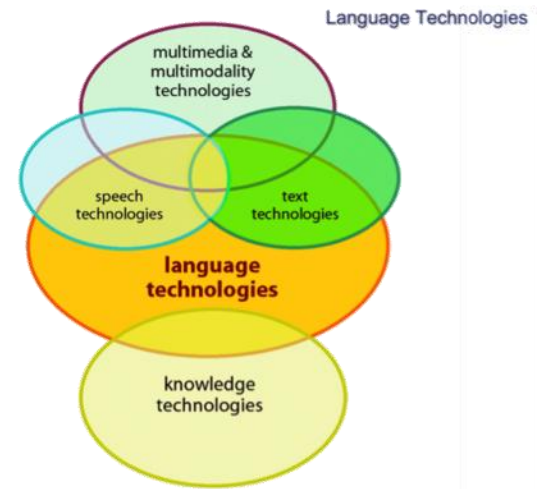
Melvyn Bragg: *The Adventure of English*, Sceptre, 2004

Bill Bryson: *Mother Tongue: The Story of the English Language*, Penguin, 2009.

# Language Technology Support for English

## Language Technologies

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



## Language Technology Application Architectures

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

- ❑ Pre-processing: cleaning up the data, removing formatting, detecting the input language, replacing contractions with their full forms, e.g., “don’t” -> “do not”, etc.
- ❑ Grammatical analysis: finding the verb and its objects, modifiers, etc.; detecting the sentence structure.
- ❑ Semantic analysis: disambiguation (Which meaning of “bank”, e.g., sloping ground by a river or financial institution, is the right one in the given context?), resolving anaphora and referring expressions like “she”, “the car”, etc.; representing the meaning of the sentence in a machine-readable way.

Task-specific modules then perform many different operations such as automatic summarization of an input text, database 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 to illustrate the complexity of language technology applications in a generally understandable way.

After the introduction of the core application areas, we will give a brief overview of the situation in LT research and education, concluding with an overview of (past) funding programs. At the end of this section, we will present an expert estimation of the situation

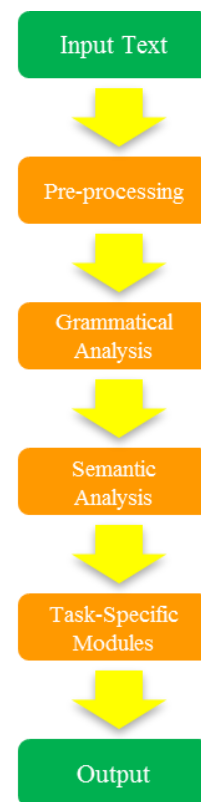


Figure 2: A Typical Text Processing Application Architecture

regarding core LT tools and resources over a number of dimensions such as availability, maturity, or quality. This table gives a good overview of the situation of LT for German.

## 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 such types of error, analysis of the context is needed in many cases. For example, the words *Eye* and *have* do not agree grammatically and so they would not, under normal circumstances, be expected to co-occur in this way. Ensuring that such grammatical mistakes are detected would require either the formulation of language-specific grammar rules, i.e., a high degree of expertise and manual labour, or the use of a statistical language model to calculate the probability of a particular word occurring along with the preceding and following words. For a statistical approach, usually based on *n-grams*, a large amount of language data (i.e. a corpus) is required to obtain sufficient statistical information.

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 incorrect usage and damage resulting from bad or poorly understood instructions, companies have begun to place an increasing focus on the quality of this technical documentation. Furthermore, as technical products came to the international market, an increasing percentage of readers were non-native English speakers. As a result, attempts were made to develop a controlled, simplified technical English that would make it easier for native and non-native readers to understand the instructional text. An example is *ASD-STE100<sup>xxiv</sup>*, originally developed for aircraft maintenance manuals, but suitable for other technical manuals. This controlled language contains a fixed basic vocabulary of approximately 1000 words, together with rules for simplifying the sentence structures. Examples of these rules include only using approved meanings for words, as specified in the dictionary (to avoid ambiguity), not writing more than 3 nouns together, always using the active voice in instruction sentences,

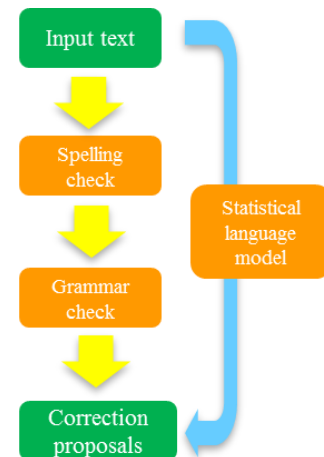


Figure 3: Language Checking (left: rule-based; right: statistical)

and ensuring that such sentences do not exceed a maximum length. The specification is maintained and kept up-to-date by the Simplified Technical English Maintenance Group (STEMG), which consists of members in several different European countries.

Advances in natural language processing have led to the development of authoring support software, which assists the writer of technical documentation to use vocabulary and sentence structures consistent with rules and terminology restrictions. The HyperSTE software<sup>xxv</sup>, developed by Tedopres International, is such an example, which is based on the *ASD-STE100* specification.

Only with the founding of the European Union did the idea that there exist many users of technical products in the world who cannot read English get attention by the manufacturers. Products that are sold in the EU have to be accompanied by technical documentation in the language of the company to which they are sold. The use of Simplified Technical English to prepare such documentation can make documentation easier to translate into other languages, and can also improve the quality of results produced by machine translation software.

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

## Web search

The search engine Google, which started in 1998, is nowadays used for almost 93% of all search queries in the UK<sup>xxvi</sup>. Since 2006, the verb *to google* has even had an entry in the Oxford English 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 facility for misspelled words and also, in 2009, incorporated basic semantic search capabilities into their algorithmic mix<sup>xxvii</sup>, which can improve search accuracy by analysing the meaning of the query terms in context. The success story of Google shows that with a large amount of data at hand and efficient techniques for indexing these data, a mainly statistically-based approach can lead to satisfactory results.

In the research labs, experiments using machine-readable thesauri and ontological language resources like WordNet have shown improvements by allowing pages to be found containing synonyms of the entered search term, e.g., the *clever search engine*<sup>xxviii</sup>. For example, if the search term *nuclear power* is entered into this engine, the search will be expanded to locate also those pages containing the terms *atomic power*, *atomic energy* or *nuclear energy*. Even more loosely related terms may also be used.

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 to determine that the user is looking for compa-

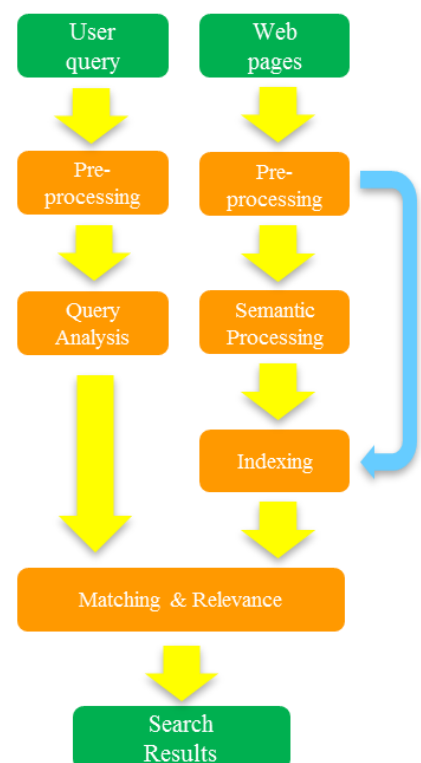


Figure 4: Web Search Architecture



nies 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 are being referenced.

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 searching for and ranking of relevant documents. In addition, to generate a list of companies, we also need to determine which particular strings of words in a document refer 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.

The first search engines for English appeared in 1993, with many having come and gone since those days. Today, apart from Google, the major players are Microsoft's Bing (accounting for approximately 4% of UK searches) and Yahoo (approximately 2% of searches in the UK, but also powered by Bing). All other engines account for less than 1% of searches. Some sites such as Dogpile provide access to meta-search engines, which fetch results from a range of different search engines. Other search engines focus on specialised topics and incorporate semantic search, an example being Yummly, which deals exclusively with recipes. Blinx is an example of a video search engine, which makes use of a unique combination of patented conceptual search, speech recognition and video analysis software to locate videos of interest to the user.

## Speech interaction

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

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.

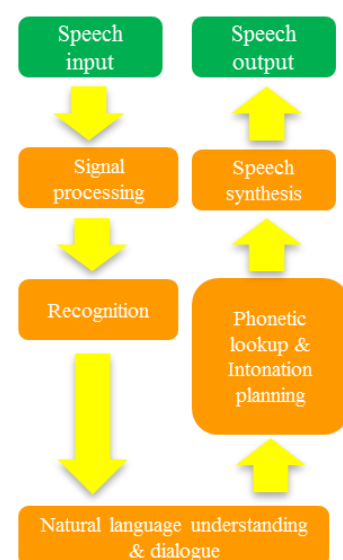


Figure 5: Simple Speech-based Dialogue Architecture

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

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

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

On the UK TTS market, Google's interest in TTS technology has been demonstrated by their recent acquisition of Phonetic Arts<sup>xxix</sup>, a company that already counted global giants such as Sony and EA Games amongst its clients. One of the selling points of Edinburgh-based CereProc is the provision of voices that have character and emotion. Roktalk is a screen reader to enhance accessibility of websites, whilst Ocean Blue Software, a digital television software provider, has recently developed a low-cost text-to-speech technology called 'Talk TV', which has the aim of making the viewing of TV more accessible to those with visual impairment. The technology has been used to create the world's first accessible technology solution designed to provide speech/talk-based TV programming guides and set up menus. The Festival Speech Synthesis System<sup>xxx</sup> is free software that has been actively under development for sev-

eral years by the University of Edinburgh, with both British and American voices, in addition to Spanish and Welsh capabilities.

Regarding dialogue management technology and know-how, markets are strongly dominated by national players, which are usually SMEs. Today's key players in the UK include Vicorp and Sabio. Rather than exclusively relying on a product business based on software licenses, these companies have positioned themselves mostly as full-service providers that offer the creation of VUIs as a system integration service. Finally, within the domain of *speech interaction*, a genuine market for the linguistic core technologies for syntactic and semantic analysis does not exist yet.

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

### Machine translation

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

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

*The policeman observed the man with the telescope.*

*The policeman observed the man with the revolver.*

One way of approaching the task is based on linguistic rules. For translations between closely related languages, a direct translation may be feasible. However, 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

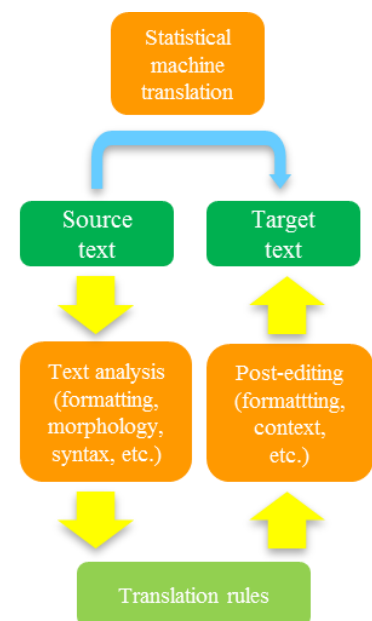


Figure 6: Machine translation (top: statistical; bottom: rule-based)

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 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.

There are several research groups in the UK and the USA active in machine translation, both in academia and industry. These include the Natural Language and Information Processing Group of the University of Cambridge, the Statistical Machine Translation Group of the University of Edinburgh, the Center for Machine Translation at the Carnegie Mellon University and the Natural Language Processing groups at both Microsoft Research and IBM Research.

SYSTRAN is one of the oldest machine translation companies, founded in 1968 in the USA and having done extensive work for the United States Department of Defense and the European Commission. The current version uses hybrid technology and offers capabilities to translate between 52 different languages. SYSTRAN is used to provide translation services on the Internet portals Yahoo, Lycos and AltaVista. Although Google originally also made use of SYSTRAN's services, they now use their own statistical-based system, which supports 57 different languages. Microsoft uses their own syntax-based statistical machine translation technology to provide translation services within their Bing search engine.

In the UK, automated translation solutions are provided by companies such as SDL, who provide a free web-based translation service in addition to commercial products. Very specialised MT systems have also been developed, e.g., the LinguaNet system, created by Cambridge-based ProLingua. This is a specially designed messaging system for cross border, mission critical operational communication by police, fire, ambulance, medical, coastguard, disaster response coordinators. It is currently used by 50 police sites in Belgium, France, the Netherlands, Spain, United Kingdom, Denmark, and Germany.

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.

## 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 stand-alone application and sometimes that of supportive, ‘under the hood’ component are text summarization and text generation. Summarization, obviously, refers to the task of making a long text short, and is offered for instance as a functionality within MS Word. It works largely on a statistical basis, by first identifying ‘important’ words in a text (that is, for example, words that are highly frequent in this text but markedly less frequent in general language use) and then determining those sentences that contain many important words. These sentences are then marked in the document, or extracted from it, and are taken to constitute the summary. In this scenario, which is by far the most popular one, summarization equals sentence extraction: the text is reduced to a subset of its sentences. All commercial summarizers make use of this idea. An alternative approach, to which some research is devoted, is to actually synthesize *new* sentences, i.e., to build a summary of sentences that need not show up in that form in the source text. This requires a certain amount of deeper understanding of the

text and therefore is much less robust. All in all, a text generator is in most cases not a stand-alone application but 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 English, question answering, information extraction, and summarization have been the subject of numerous open competitions since the 1990s, primarily organized by DARPA/NIST in the United States, which have significantly improved the state of the art. For example, the annual TREC (Text REtrieval Conference) series included a question-answering track between 1999 and 2007. Recently, freely accessible tools have been developed that reason and compute the answers. These include True Knowledge, developed in the UK, and Wolfram Alpha, developed in the USA. Question-answering systems in more specialist domains have also begun to emerge, such as the EAGLi system for questions answering in the Genomics literature, developed at the University of Applied Sciences, Geneva.

Information Extraction research was boosted by both the series of MUCs (Message Understanding Conferences), running from 1987-1998, and subsequently by the Automatic Content Extraction (ACE) program, running from 1999 to 2008. Domain-specific challenges such as BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology), of which the most recent was held in 2010, have helped to further research into Information Extraction from more specialized types of text. Evaluation of text summarization systems was carried out as part of the Document Understanding Conferences (DUC) from 2001-2007, and more recently as one of the tracks in the Text Analysis Conferences (TAC). Web-based tools such as Ultimate Research Assistant and iResearch Reporter can produce summary reports of retrieved search results.

## Language Technology in Education

In the UK, a large number of universities have well-established research groups that are active in the field of language technology or computational linguistics. These are complemented by many other groups in English speaking countries, most notably the USA, Australia and Ireland. These groups are most often part of either computer science or linguistics departments. The University of Manchester hosts the National Centre for Text Mining (NaCTeM), which is the world's first publicly funded text mining centre, providing text mining services to both academic institutions and industrial organisations. Over the past few years, there has been an increasing interest in tools and resources dealing with specialist domains such as biomedicine, molecular biology and chemistry.

In terms of teaching in the UK, courses with a large element of natural language processing or computational linguistics are rare, and are normally only offered at the masters level. Examples include the MSc in Speech and Language Processing and the MSc in Cognitive Science, offered at the University of Edinburgh. A greater number of universities offer course modules in NLP to students of more general degree programs. Examples include Birmingham, Cambridge, Manchester, and Leeds.

## Language Technology Programs

The first working demonstration of an LT system took place in the 1950s. This system constituted a Russian – English Machine

Translation (MT) system, developed by IBM and Georgetown University. The company SYSTRAN, which was founded in 1968, had the original aim of processing the same language pair for the United States Airforce. SYSTRAN still exists today, as described in the *Machine translation* section above.

An early LT program, EUROTRA, was an ambitious Machine Translation (MT) project inspired by the modest success of SYSTRAN, and established and funded by the European Commission from the late 1970s until 1994. The project was motivated by one of the founding principles of the EU: that all citizens had the right to read any and all proceedings of the Commission in their own language. A large network of European computational linguists embarked upon the Eurotra project with the hope of creating a state-of-the-art MT system for the then seven, later nine, official languages of the European Community. However, as time passed, expectations became tempered; "Fully Automatic High Quality Translation" was not a reasonably attainable goal. The true character of Eurotra was eventually acknowledged to be in fact pre-competitive research rather than prototype development. While Eurotra never delivered a "working" MT system, the project made a far-reaching long-term impact on the nascent language industries in European member states.

The Alvey Programme was the dominating focus of Information Technology research in the UK between 1983 and 1988. Amongst the areas of focus was Man Machine Interaction. The programme funded three projects at the Universities of Cambridge, Edinburgh and Lancaster to provide tools for use in natural language processing research. The tools, a morphological analyser, parsers, a grammar and lexicon were usable individually as well as together - integrated by a grammar development environment - forming a complete system for the morphological, syntactic and semantic analysis of a considerable subset of English.

The creation of the British National Corpus (BNC) was a major project that took place between 1991 and 1994. The corpus constitutes a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century. The corpus is encoded according to the Guidelines of the Text Encoding Initiative (TEI) to represent both the output from CLAWS (automatic part-of-speech tagger) and a variety of other structural properties of texts (e.g. headings, paragraphs, lists etc.). An XML version of the corpus was released in 2007.

Corpora of other varieties of English are also being collected. The International Corpus of English (ICE), whose collection began in 1990, involves 23 research teams around the world, who are preparing electronic corpora of their own national or regional variety of English. Each team is producing a corpus consisting of one million words of spoken and written English produced after 1989. The Corpus of Contemporary American English (COCA) consists of 425 million words, equally divided among spoken, fiction, popular magazines, newspapers, and academic texts, consisting of 20 million words each year from 1990-2011.

AKT (2000 – 2007), was a multi-million pound collaboration between 5 UK universities with the aim of enhancing information and knowledge management in the age of the World Wide Web. The team of 119 staff was interdisciplinary, involving leading figures in

the worlds of multimedia, natural language and computational linguistics, agents, artificial intelligence, formal methods, machine learning and e-science. The research conducted on the project formed an important contribution to the semantic web, in which the use of LT technologies played a central role. The AKT collaboration was a significant success in terms of papers published, grants awarded (36 other projects), students trained and international impact. It was rated as “outstanding” by the review panel. The collaboration placed major importance on making links with industrial partners, and finally it led to the founding of a number of spin-off companies. A follow-up project, EnAKTing the Unbounded Data Web: Challenges in Web Science, is currently ongoing.

Since many LT applications make use of similar sets of processing components, such as tokenizers, taggers, parsers, named entity recognisers, etc., the speed with which new applications can be developed can be greatly increased if such processing components can be reused and repurposed in flexible ways to create a range of different LT applications. Two systems which support the user in creating new applications from existing libraries of processing components are the University of Sheffield’s GATE system, which has been under development for over 15 years, and the more recent U-Compare system, which was developed as part of a collaboration between the Universities of Tokyo, Manchester and Colorado. Whilst current components in U-Compare mainly deal with English, the library will be extended as part of META-NET to cover a number of different European languages.

### Status of Tools and Resources for English

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
<b>Language Technology (Tools, Technologies, Applications)</b>							
Tokenization, Morphology (tokenization, POS tagging, morphological analysis/generation)	6	6	5	4	5	3	5
Parsing (shallow or deep syntactic analysis)	5	4	5	4	4	3	3
Sentence Semantics (WSD, argument structure, semantic roles)	4	4	4	3	3	2	3
Text Semantics (coreference resolution, context, pragmatics, inference)	3	3	2	1	2	1	1
Advanced Discourse Processing (text structure, coherence, rhetorical structure/RST, argumentative zoning, argumentation, text patterns, text types etc.)	2	2	2	2	2	2	1
Information Retrieval (text indexing, multimedia IR, crosslingual IR)	5	5	4	5	4	4	4
Information Extraction (named entity recognition, event/relation extraction, opinion/sentiment recognition, text mining/analytics)	5	5	3	4	5	3	3



	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Generation (sentence generation, report generation, text generation)	2	2	2	2	2	2	2
Summarization, Question Answering, advanced Information Access Technologies	3	3	2	2	2	2	2
Machine Translation	5	4	2	3	5	2	2
Speech Recognition	5	3	4	4	4	2	3
Speech Synthesis	5	3	4	5	4	2	3
Dialogue Management (dialogue capabilities and user modelling)	3	2	4	3	4	2	5
<b>Language Resources (Resources, Data, Knowledge Bases)</b>							
Reference Corpora	5	5	5	4	5	3	3
Syntax-Corpora (treebanks, dependency banks)	5	2	6	4	5	2	5
Semantics-Corpora	3	4	3	3	3	2	2
Discourse-Corpora	3	3	3	3	2	2	2
Parallel Corpora, Translation Memories	4	4	4	4	3	3	3
Speech-Corpora (raw speech data, labelled/annotated speech data, speech dialogue data)	5	2	6	5	5	3	3
Multimedia and multimodal data (text data combined with audio/video)	2	1	2	1	1	2	1
Language Models	6	4	5	5	5	3	6
Lexicons, Terminologies	5	6	5	4	4	5	5
Grammars	3	2	3	3	2	4	1
Thesauri, WordNets	3	6	4	5	5	4	4
Ontological Resources for World Knowledge (e.g. upper models, Linked Data)	4	4	4	4	3	3	3

## Conclusions

Since English is the language with the longest history of LT research, a wide range of high quality tools and resources exist, often covering a greater number of areas and being available in larger numbers than for other European languages. However, the overall situation regarding language technology support for English still



needs to be approached with a certain amount of caution; in some areas, it is not as well-developed as may be expected, and is still a long way from reaching the necessary state to support a truly multilingual knowledge society.

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 English, key results regarding technologies and resources include the following:

- ❑ No single category of technology or resources has consistently high scores across all criteria being evaluated.
- ❑ Generally, quantity, quality and availability can only be guaranteed for tools and resources dealing with more basic levels of linguistic processing.
- ❑ Higher levels of linguistic processing (i.e., semantics and discourse) still present considerable challenges. The lower number of corpora annotated with these levels of information could be a factor limiting the advancement of these technologies, since the development of such technologies is more difficult if the amount of data on which they can be trained is limited.
- ❑ Sustainability is, in general, a major area of concern. Even if high quality technologies and resources exist, major efforts may still be required to ensure that they are kept up-to-date and can easily be integrated into other systems. There is also often a lack of rigorous software testing/engineering principles applied to tools. The availability of the high-performance Lucene search engine for Information Retrieval, and the high quality test suites for grammar engineering, make these two areas notable exceptions.
- ❑ In general, tools that work well on a particular type of text may require considerable work to allow them to be applied to new text domains. Resources such as annotated corpora are also normally domain-specific, and creating such corpora for new domains generally requires a large amount of manual work.
- ❑ For all technologies and tools, there are examples that are available free of charge. However, the number of such tools and resources varies greatly according to category. In some cases, quality comes at a price. For example, in the case of syntactic corpora, there is little to rival the Penn TreeBank, which is only available for a fee. In other cases, even large corpora are available free of charge, e.g. Google's n-gram corpus for statistical language modelling, which was created from 1 trillion word tokens of text from publicly accessible Web pages.
- ❑ Some broad areas, such as Information Extraction, consist of a number of component technologies. Whilst some of these technologies (e.g. named entity tagging), are quite mature and can produce high quality results, others, such as event/relation extraction are more complex and still require improvement. The scores awarded attempt to balance the different stages of development of these technologies.

It is without doubt that there exist extremely strong foundations on which the already thriving language technology landscape for English can continue to grow and prosper. However, it is important to emphasize that many aspects of language technology have still yet to be solved. In certain cases, some of these problems concern the



need to focus greater research efforts on some of the more complex areas of LT, including advanced discourse processing and language generation. However, some more general issues, including problems of sustainability and adaptability, which are common across many types of tools and resources, are in urgent need of more focussed strategies.

## 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.

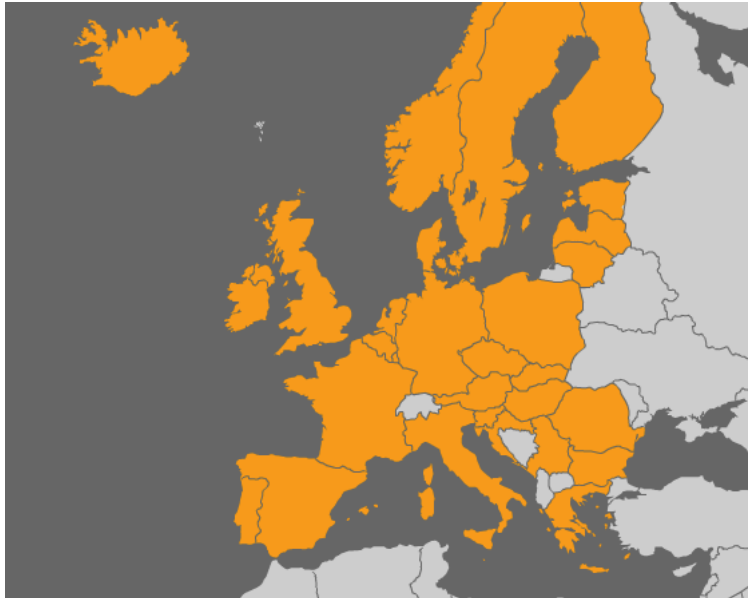


Figure 1: Countries Represented in META-NET

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



*The Multilingual Europe Technology Alliance (META)*

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 FLReNet Forum (Spain), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Malta), EAMT 2010 (France) and ICT 2010 (Belgium) centred on public outreach. According to initial estimates, META-NET has already contacted more than 2,500 LT professionals to develop its goals and visions with them. At the META-FORUM 2010 event in Brussels, META-NET communicated the initial results of its vision building process to more than 250 participants. In a series of interactive sessions, the participants provided feedback on the visions presented by the network.

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

**META-RESEARCH** builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, this activity wants to bring more semantics into machine translation (MT), optimise the division of labour in hybrid MT, exploit context when computing automatic translations and prepare an empirical base for MT. META-RESEARCH is working with other fields and disciplines, such as machine learning and the

Semantic Web community. META-RESEARCH focuses on collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community. This activity has already clearly identified aspects of MT where semantics can impact current best practices. In addition, the activity has created recommendations on how to approach the problem of integrating semantic information in MT. META-RESEARCH is also finalising a new language resource for MT, the Annotated Hybrid Sample MT Corpus, which provides data for English-German, English-Spanish and English-Czech language pairs. META-RESEARCH has also developed software that collects multilingual corpora that are hidden on the web.

## Member Organisations

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

Country	Organisation	Participant(s)
Austria	University of Vienna	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernelle
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	Eiríkur 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 Pezik
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

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



## References

---

- <sup>i</sup> 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](http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf)).
- <sup>ii</sup> 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](http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf)).
- <sup>iii</sup> UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (<http://unesdoc.unesco.org/images/0015/001503/150335e.pdf>).
- <sup>iv</sup> 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>).
- <sup>v</sup> <http://www.efnil.org/documents/language-legislation-version-2007/united-kingdom/english>
- <sup>vi</sup> <http://www.legislation.gov.uk/mwa/2011/1/section/1/enacted>
- <sup>vii</sup> Laubach, F.C. (1996). *Let's Reform Spelling - Why and How*. NY: New Readers Press.
- <sup>viii</sup> <http://www.oed.com/public/latest/latest-update/>
- <sup>ix</sup> [http://ec.europa.eu/public\\_opinion/archives/ebs/ebs\\_243\\_sum\\_en.pdf](http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_sum_en.pdf)
- <sup>x</sup> <http://www.le.ac.uk/engassoc/>
- <sup>xi</sup> <http://www.ccue.ac.uk/>
- <sup>xii</sup> <http://www.nate.org.uk/>
- <sup>xiii</sup> <http://www.essenglish.org/>
- <sup>xiv</sup> <http://www.queens-english-society.com/>
- <sup>xv</sup> [http://curriculum.qcda.gov.uk/uploads/English%201999%20programme%20of%20study\\_tcm8-12054\\_tcm8-16038.pdf](http://curriculum.qcda.gov.uk/uploads/English%201999%20programme%20of%20study_tcm8-12054_tcm8-16038.pdf)
- <sup>xvi</sup> [http://www.cambridgeassessment.org.uk/ca/digitalAssets/113995\\_Stats\\_report\\_5\\_-\\_A\\_level\\_uptake\\_2006.pdf](http://www.cambridgeassessment.org.uk/ca/digitalAssets/113995_Stats_report_5_-_A_level_uptake_2006.pdf)
- <sup>xvii</sup> <http://www.pisa.oecd.org/>
- <sup>xviii</sup> <http://www.efnil.org/documents/language-legislation-version-2007/united-kingdom/english>
- <sup>xix</sup> [http://eacea.ec.europa.eu/about/eurydice/documents/KDL2008\\_EN.pdf](http://eacea.ec.europa.eu/about/eurydice/documents/KDL2008_EN.pdf)
- <sup>xx</sup> Swales, J. M. (1997), *English as Tyrannosaurus rex*. *World Englishes*, 16: 373–382
- <sup>xxi</sup> <http://www.statistics.gov.uk/ci/nugget.asp?id=8>
- <sup>xxii</sup> <http://www.internetworldstats.com/stats7.htm>
- <sup>xxiii</sup> <http://www.denic.de/hintergrund/statistiken/internationale-domainstatistik.html>
- <sup>xxiv</sup> <http://www.asd-ste100.org/>
- <sup>xxv</sup> <http://www.simplifiedenglish.net/HyperSTE-Software/>
- <sup>xxvi</sup> [http://gs.statcounter.com/#search\\_engine-GB-monthly-201010-201012](http://gs.statcounter.com/#search_engine-GB-monthly-201010-201012)

xxvii

[http://www.pcworld.com/businesscenter/article/161869/google\\_rolls\\_out\\_semantic\\_search\\_capabilities.html](http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html)

xxviii <http://wdok.cs.uni-magdeburg.de/clever-search/>

xxix <http://googleblog.blogspot.com/2010/12/can-we-talk-better-speech-technology.html>

xxx <http://www.cstr.ed.ac.uk/projects/festival/>