University of Zurich
MA Multilingual Text Analysis

Report Study Week MLTA
03 to 09 June 2012
in Donostia-San Sebastián, Spain

The Benefits of Language Technology for Multilinguality – The Benefits of Multilinguality for Language Technology. The Case of the Basque Country

Esther Germann (MLTA)
Franziska Tobler (MLTA)
Jeanette Isele (MLTA)
Katrin Rettich (MLTA)
Lenz Furrer (MA Computerlinguistik)
Manuela Weibel (MLTA)
Mirjam Marti (MLTA)
Susanna Tron (MLTA)

English Version
2012-08-23

1 A study week in the Basque Country – An introduction

The international study week of the specialized masters program 'Multilingual Text Analysis' (MLTA) focused on the question how language technology is able to benefit from a multilingual environment and how multilingual societies can gain from current language technology. In general, the master program addresses technological editing and processing of multilingual documents this led to an investigation of the complex question concerning the connection between technology and language diversity in the multilingual Basque country in Spain. The Spanish part of the Basque Country is situated on the Atlantic coast right on the border to France. It is composed of the four provinces Gipuzkoa, Biskaya, Álava and Navarra. There are two official languages in the Spanish Basque Country. Besides Spanish, the regional language Basque is used as an official language as well. A linguistic relationship between Basque and another language was never established. The Basque language used to be widely repressed during the time of the Spanish dictatorial regime, however, since 1978, it has received profound stabilization in society and today, possesses around 500.000 speakers in Spain[1], who converse in seven separate Basque dialects. The standardization and development of a Basque written language named *Euskara Batua* led to Basque being represented in education, industry, administration, research and media of the Basque Country.

The eight students participating in the study week from the 4th until the 8th of June, 2012 visited together with MLTA program manager Martin Volk and program coordinator Jeannette Roth a variety of institutions and companies of the Basque country. This report gives an insight to these visits and the addressed topics and projects regarding the question of the influence between multilingualism and language technology.

1 http://de.wikipedia.org/wiki/Baskische_Sprache

2 Vicomtech in Donostia-San Sebastián (2012-06-04)

2.2 Basque

At beginning of the MLTA students visit at Vicomtech, the structure of the Basque language (Basque *euskara*) was presented by a (female) linguist. It is a genetically isolated language which is taken as pre-indogermanic. Basque possesses a highly agglutinative morphology. Another characteristic of Basque is its polypersonalism. For instance, in one verb form, several persons can be marked by affixation. In addition, Basque has 16 case markers which are implemented by suffixes on nouns. This leads to a very complex morphology of the verbs. A unique characteristic of Basque is its ergativity. If a sentence contains an intransitive verb, the subject is unmarked and in the absolutive case. However, if it the sentence is transitive, the object is in the absolutive case and unmarked whereas the subject is in the ergative case, and thus, unmarked.

2.3 Vicomtech

Vicomtech[2] is a center for applied science in the domains of visual interaction and communication technology which is located within the Spanish-Basque part of San Sebastián. The enterprise develops technologies for companies and institutions and as a result collaborates closely with industrial partners as well as local institutions such as the IXA-Group and Elhuyar. What is more, Vicomtech is part of several European projects such as the SUMAT-project which will be outlined in the following. The enterprise is also part of the technological alliance IK4, which unites the eight leading Basque centers for technology. Vicomtech has about 100 national and international employees and with over 90 annual projects, tries to build a bridge between local enterprises and the international economy. With Vicomtech's aid local businesses are given a chance to profit from international and current technological research.

2 www.vicomtech.org, retrieved June 23, 2012

2.4 Human speech and language technologies at Vicomtech

Since 2011, Vicomtech has a division for human speech and language technology which is run by Dr. Arantza del Pozo Echezarreta. She was responsible for hosting the MLTA students' visit. The division was created in response to a need for language technology components in the research on audiovisual applications. Approximately ten researchers concern themselves with the development of natural language processing applications which can be used later by language technology companies in the interdisciplinary field of the intersection between man and machine and multilingual knowledge management. The division covers three foci: speech processing, natural language processing and the area of dialog systems.

In the course of the MLTA student's half-day visit at Vicomtech, different projects were introduced by their researchers. In the following, the finalized project BerbaTek, the ongoing projects SUMAT and CAPER as well as the future project OpeNER will be presented.

2.5 Projects at Vicomtech

BerbaTek

At the end of 2011, the research project BerbaTek[3] was completed. It was concerned with the development of language and multimedia technology. The aim of the project was to improve the contestability of Basque companies in the language technology sector. For this purpose, applications in the range of translation, language learning and content management were explored. The Vicomtech team developed in the scope of BerbaTek an audiovisual system for question-answering in Basque which is illustrated with the aid of an avatar. In addition, some work on the automatic synchronization of films was done as well. However, the issue of voice transformation in order to use one voice actor for several roles in one film remained unresolved. This application did not meet the standards of the industry and remains an open research field. On top of this the researchers built a parallel corpus (Spanish-Basque) which was based on the bilingual legal and administrative texts of the Basque Country and which was among other things used for the development of a Basque-Spanish machine translation system.

3 http://www.vicomtech.es/ingles/html/proyectos/index_proyecto158.html

SUMAT

The EU project SUMAT[4] works on an online-service for subtitling by machine translation (*An Online Service for Subtitling by Machine Translation*). A project partner, besides Vicomtech and other European firms, is TextShuttle a University of Zurich spin-off. The translation service for subtitles is being developed for nine different European languages. In the end, users will be able to choose from 14 possible language pairs for their translations. Since more and more digital multimedia such as films is offered on the European market, a fast and high quality opportunity is called for which allows subtitling in order to enhance comprehensibility. The SUMAT-project views such multilingual demands as an opportunity to make statistical machine translation in collaboration with human post-editors available for the subtitling industry. The online-service is supposed to be used efficiently for the creation of subtitles by free-lance translators as well as by subtitling firms and thus, help to enhance the industry's productivity.

CAPER

Besides SUMAT Vicomtech is involved in another EU project called CAPER[5], which researches the joint acquisition, processing, exploitation and reporting of information in order to prevent organized crime (c*ollaborative information acquisition, processing, exploitation and reporting for the prevention of organized crime*). The aim of the project is to build a platform on which public and private information on organized crime in Europe can be exchanged. In order to obtain information multilingual, audio-visual content and biometric information are analyzed. For this purpose, the technology of data mining is used among other things. The project uses techniques such as named entity recognition to identify persons involved in crimes and to pinpoint who is or was involved in which action. For processing the names, the team built a NLP-pipeline. In addition, the collected data are standardized for replacement. The project aims to make the final results user-friendly so that the information platform can be used by Law Enforcement Agencies (LEAs) in Europe. A challenge for the scientists is to make multilingual and multi-modal contents compatible with the aid of semantic analysis techniques and to develop appropriate tools for the end user.

4 http://www.sumat-project.eu/node/2
5 http://www.fp7-caper.eu/

OpeNER

As of July 2012, Vicomtech scientists will be working on the EU project OpeNER[6]. The aim of the project is to identify opinions about tourist locations of all kinds by using named entity recognition and sentiment analysis. Consumer reviews about hotels, cities, restaurants, sights, and other tourist places will be collect and supplied to tourist agencies as well as to the enterprise (hotel, restaurant, etc.) in question. A major challenge is the multilingual study of the selected languages Spanish, English, French, German, Dutch and Italian. Tripadvisor was obtained as a partner of the project which will probably provide many of the place, hotel and restaurant names for analysis. Moreover, all of the techniques and tools developed in the project will be available as open source tools.

2.6 Conclusion

The Basque enterprise Vicomtech in its division *Human Speech and Language Technologies* works at many different research projects and covers many areas of current language technology. But not only the variety of topics, but also the collaboration with several other firms, institutions and EU projects and the associated internationalization render Vicomtech a convincing and interesting enterprise. A positive feature of the division is the strong integration of PhD students and university graduates. Moreover, Vicomtech plays a big role in the development of language technologies for the Basque language. Applications such as the automatic generation of subtitles or automatic synchronization contribute to the stabilization of the Basque language. The multilingualism of the Basque country and the promotion of the Basque language, clearly benefit from Vicomtech's developments in the field of language technology.

6 http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_LANG=EN&PJ_RCN=12917615&pid=3&q=F0CBA7CEEF53305FB0AC9C3F958A6CCF&type=rap

3 Mondragon Lingua in Arrasate-Mondragón (2012-06-04)

3.1 Mondragón and the *Mondragon Corporation Cooperativa*

Mondragón (Basque *Arrasate*) is a small village (just under 20.000 inhabitants) in the Basque province Gipuzkoa which is located in the direction of the inland about 70km off San Sebastián.

In 1943, the priest José María Arizmendiarrieta laid the foundation for the small town's big success by establishing a democratically organized university of applied science. When the university had produced a number of successful cooperatives, in 1956, *Mondragón Corporación Cooperativa* (MCC) was founded. Until today, its main aim is to systematically create employment opportunities. Currently, MCC consists of more than 100 enterprises in three very different sectors: the financial sector (e.g. banking institutions, insurances), the industrial sector (e.g. automobile industry) and the retail industry (e.g. supermarket chains). In 1997, MCC even founded its own private university in Arrasate-Mondragón, which carries the Basque name *Mondragon Unibersitatea*. Today, the university is made up of four departments (engineering, gastronomy, humanities and educational sciences) and includes around 4000 students. An important advantage of the university is its closeness to the employment market. The students benefit from the university's market-oriented study programs.

3.2 Mondragon Lingua

In 1973, in the service department within the industrial sector of the MCC, the speech center CIM (*Centro de Idiomas Mondragón*) was established. Today it is called *Mondragon Lingua*. Subsequently the corporation has built up a reputation as a language school which as early as the beginning of the 1980s used information technology to support its language courses. That is why it is considered a pioneer in the use of computers for language teaching. Currently the company has five centers in the Basque Country: in Bilbo-Bilbao, Donostia-San Sebastián, Arrasate-Mondragón, Gasteiz-Vitoria and Oñati-Oñate. In addition to language courses, Mondragon Lingua offers translation and interpreting services. At the present time, over 100 highly qualified translators and interpreters who are able to translate into more than 40 languages work for Mondragon Lingua.

3.3 Mondragon Lingua and Machine Translation

Currently Mondragon Lingua is developing a project which aims to integrate machine translation for the language pair English-Spanish into the conventional workflow of human translation. The project heads are Ane Ruiz de Zarate and Manuel de la Pascua. According to de la Pascua, the training data consists of “several 10 million words” and they will use Moses as their machine translation system. Since the project is still in it early stages, no definite statements about the specific approach or the success of the project could be made.

During the project heads' presentation of the project it became obvious, that the integration of machine translation proceeds reluctantly still (quote from the PowerPoint presentation: “MT – because we have no other option“). Nevertheless we draw a pivotal conclusion from the visit at Mondragon Lingua: the translation industry increasingly recognizes the need for machine translation and little by little begins to participate.

4 The IXA Research Group in Donostia-San Sebastián (2012-06-05)

The University of the Basque Country (Spanish *Universidad del País Vasco*, Basque *Euskal Herriko Unibersitatea*) has about 45 000 students and 5000 members of academic staff.[7] Roughly 100 departments are spread across four locations in Bilbao, Vitoria-Gasteiz, Eibar and San Sebastián. The University is bilingual: About 35 % of the students study in Basque.

The *IXA Research Group*[8] belongs to the Department of Computer Languages and Systems and is located in San Sebastián. The group was founded in 1988 with the goal of developing basic language technology resources the Basque language, though their research is nowadays more generally described as *Natural Language Processing.* Currently the group employs an interdisciplinary team consisting of 33 computer scientists, 11 linguists, and 3 technical assistants. 28 PhD teachers of the university are connected with IXA through various projects. 22 doctoral theses (including 6 external ones) are being written here and 8 students are completing their studies.

IXA collaborates with organizations such as Vicomtech or Elhuyar Foundation, as well as with universities in Barcelona, Amsterdam, Rome and Darmstadt.

4.1 Research Activity

In an introductory presentation, Arantza Díaz de Ilarraza gave us an overview of the activities of IXA. The research group is concerned with general problems of language processing, especially with regard to the Basque language. They are strategically promoting Basque language technology by creating linguistic (theoretical) foundations and resources (such as corpora) which are necessary for automatic processing. Tools for further development or even complete applications are also part of their program. The list of manufactured products is considerable and covers areas such as information retrieval and information extraction, machine translation, language learning software, morphology, (morpho-)syntax, and lexicography / semantics.

One of IXA's flagships is the Basque spell checker *Xuxen* (meaning about 'exactly correct'), which is freely available for word processing programs (MS Office, OpenOffice), the Mozilla products (browser, e-mail client) and as an online application.[9] The importance of a Basque spell checker is to be emphasized: Many speakers of the Basque language are very inconfident about orthography, since there had been no standardized written language at all until 1966[10] and only around 1980 the Basque language started to be used in the educational system.

IXA was involved in the creation of several extensive digital resources. In order to build the spell checker, the EDBL *(Euskararen Datu-Base Lexikala, '*Lexical Database for Basque') was created, which was later also used for other applications (morphology analysis, lemmatization). It covers about 100 000 lemmas with morphological information. Another lexical resource is the Basque WordNet *BasWN,* a knowledge base which arranges the meanings of words in structured lexical-

7 www.ehu.es

8 *ixa* (pronounciation: [iʃa]) is the Basque name of the letter X.

9 www.xuxen.com (in Spanish and Basque only)

10 see the handout IXA Research Group of 06/05/2012, Slide 9

semantic relations. It contains nearly 60 000 words with a total of 90 000 meanings (synsets). Also, two Basque corpora have been created: EPEC *(Euskararen Prozesamendurako Erreferentzia Corpusa*, 'reference corpus for the processing of Basque') is a corpus of 300 000 words containing manual annotation for morphology, phrases, syntactic dependencies, semantic roles and synsets from BasWN. The newer, only morphosyntactically annotated corpus ZTC *(eta Zientzia Teknologiaren Corpusa,* 'science and technology corpus') is considerably larger with its 10 million words; around a fifth of the automatically generated annotations were corrected by hand. A freely accessible online interface[11] allows complex queries to the ZTC.

The size of the corpus led to the question, how such an amount of scientific texts could be made available in Basque, and if there is actually publishing in Basque. The majority of technological or other internationally interesting publications are mainly written in English, those of national interest, however, often in Spanish. There are also publications that reach a Basque speaking audience. The university, which offers a considerable part of the courses in Spanish and Basque in parallel, has thus also a high demand for Basque literature. However, only a small part was available in digital form, a lot of printed material had to be digitized for the corpus.

4.2 KYOTO – Knowledge Yielding Ontologies for Transition-based Organization

Aitor Soroa presented us the *KYOTO* Project, a multi-lingual environmental text-mining system. The project is funded by the EU and aims at creating an open, freely accessible platform for environmental knowledge. The project is led by Professor Piek Vossen of the University of Amsterdam and includes research groups in Italy, Spain, the Czech Republic, the Republic of China, Japan and Germany.

The idea is that the text-mining system automatically finds events (facts) in texts written in different languages and makes them accessible across languages and cultures.

It supports the languages Spanish, English, Dutch, Italian, Chinese and Japanese. WordNet functions as a bridge between the languages. The non-English wordnets (see EuroWordNet[12]) are structurally based on the English WordNet and provide links to the latter. Thus, the different expressions in the texts refer to their corresponding concepts.

In a first step, the system creates a terminology database. The so-called 'TyBots' *(Term Yielding Bots)* extract terms with the help of Wikipedia, connect them with concepts and build up an ontology.

The actual knowledge extraction is carried out with several hundred hand-written rules. The so-called 'KyBots' *(Knowledge Yielding Bots)* generate an output template for each sentence, as soon as a rule matches. The rules have each a morphosyntactic component and a semantic condition, both of which must be met. This is the most difficult task, since the wide range of variation in natural language must be taken into account.

The extracted knowledge is finally reported on three levels: lexically (terms), WordNet (concepts)

11 www.ztcorpusa.net/cgi-bin/kontsulta.py (in Basque only)
12 www.illc.uva.nl/EuroWordNet/

and in the ontology. The extracted facts are stored in a specific XML format *(KAF,* Kyoto Annotation Framework), which is suitable for linguistic and semantic processing and at the same time flexible and extensible.

4.3 Opener – Open Polarity Enhanced named entity recognition

The project *OPENER* was presented to us by Rodrigo Agerri. The project is a collaboration with Vicomtech and is described in more detail in Section 2.5.

4.4 Machine Translation

Gorka Labaka informed us on behalf of the IXA group's activities in machine translation. The system *Matxin,* that some of us already knew, was developed here. It is corpus- and rule-based and supports the language pairs Spanish-Basque and English-Basque.

The new hybrid MT system *SMatxinT* shall now bring the benefits of statistical MT into the translations. The advantages of Statistical MT over rule-based MT are usually seen in better lexical coverage and selection, and better handling of unknown structures.

Compared to statistical MT systems for most Western European languages, the Basque language, being agglutinative and thus extremely rich in morphology, offers additional difficulties. Due to the rich inflection, Basque has a very high types-per-token ratio, which leads to sparseness in the training data. To make matters worse, the availability of parallel data in Basque is rather limited. At the same time, Basque usually has less tokens per sentence than, for example, the Spanish translation, leading to frequent one-to-many alignments. This tends to affect the translation accuracy, since there is less generalization.

This problem can be decreased by using a morphology system to segment the words into stems and affixes or clusters of affixes. With the right choice of split points, Basque verbal affixes, for example, can be aligned to Spanish auxiliary verbs, which increases the proportion of one-to-one alignments and reduces the number of types. Thus, even word forms that had not been seen during training might be translated correctly.

The hybrid system SMatxinT uses a rule-based subsystem as the skeleton for the translation. The structure of the target sentence is determined by syntax rules. The statistical subsystem then computes several translation candidates for phrases or other sentence fragments alternative to those of the rule-based system, among which the best is selected.

4.5. Semantic Analysis

The afternoon brought a short introduction into the primary and most relevant databases used in IXA's projects. In the following, a few of the discussed resources will be presented briefly.

General Inquirer (1966)

On its homepage the *General-Inquirer*-Projekt[13] is described as a computational approach **to** content analysis of texts as well as a mapping tool. Rodrigo Aguerri explained that the projects is a resource for opinions in which words are categorized into positive or negative values. In practice this means that each text file is mapped according to categories. The categories, of which there is

13 http://www.wjh.harvard.edu/~inquirer

a total of 182, consist of a list of words and meanings. The project states that the category “negative” is the biggest with 2291 entries. In addition, users have the option of creating their own categories. However, the tool is monolingual, i.e. it is only available in English. It is one of the resources which were used in IXA's basic research.[14]

EuroWordNet

EuroWordNet[15] professed the aim to establish a multilingual database with WordNets in several European languages. It was finished in 1999. The languages which were used are Dutch, Italian, Spanish, German, French, Czech and Estonian. Similar to a variety of WordNet projects, *EuroWordNet* is based on the Princeton WordNet for American English. WordNets are structured in such a way as to allow categories of synonyms (synsets) to be linked to categories in other languages. However, each language has its own specific system for lexicalization. The conjunction of the different language lexica is made by the so-called Inter-Lingual-Index (by means of synonyms, semantic relationships). In *EuroWordNet* an additional top-ontology exists comprising of 63 semantic distinctions. This ontology is shared by all the languages. *EuroWordNet* was presented as an example for a basic WordNet which served IXA as inspiration and starting point for their own projects such as the Basque WordNet.

14 http://www.wjh.harvard.edu/~inquirer/3JMoreInfo.html
15 http://www.illc.uva.nl/EuroWordNet/

SentiWordNet

There are different versions of *SentiWordNet*[16] (1.0 and 3.0). Essentially it is a lexical resource for opinion mining. SentiWordNet ascribes each synset of a WordNet three sentiment values: positivity, negativity and objectivity. For IXA's research SentiWordNet 1.0, which was established in 2006, is particularly relevant.

WordNet-Affect

WordNet-Affect[17] is an expansion of another WordNet project: *WordNet Domains*. As the name implies, this WordNet covers an additional category of synsets which serves to depict affective concepts and words. These WordNet-synsets receive affective designations (*a-labels*) which are used to designate emotionally oriented concepts.

Q-WordNet

Q-WordNet[18] is one of the (ongoing) projects of Rodrigo Agerri which deal with sentiment analysis. It is available as download on his personal homepage. This WordNet is a lexical resource consisting of WordNet meanings which were automatically assorted to a positive or negative polarity. Polarity classification is mainly employed in attitude research such as the *OpeNER* project. *Q-WordNet* was inspired by *SentiWordNet*. In contrast to *SentiWordNet*, the linguistic information was annotated automatically as well as manually. For Rodrigo Agerri and his team this additional effort meant more work but also an important qualitative improvement of the material.

4.6 Conclusion

The IXA research group highly values multilingualism. Since it is the university of the Basque Country, Basque is a natural focus of education and research. Basque is promoted actively not just in educational activities but also in research projects such as the spell checking program Xuxen and the creation of digital resources (EDBL, BasWN, EPEC and ZTC). Through the use of the language pairs Spanish-Basque and English-Basque the machine translation system *SMatxinT* contributes to bridging language barriers. Ongoing projects such as *Kyoto* and *OpeNER* extend IXA's multilingual orientation beyond West European languages.

16 http://sentiwordnet.isti.cnr.it/
17 http://wndomains.fbk.eu/wnaffect.html
18 http://www.rodrigoagerri.net/sentiment-analysis

5 The San Telmo Museum in Donostia-San Sebastián (2012-06-06)

The San Telmo Museum in San Sebastián was opened in 1902 and is the oldest museum in the Basque Country. First, it included the disciplines of art history and archeology; a growing demand and necessity led to a fourth field, ethnography being added.[19] The main goal of San Telmo is to promote the historical and socio-cultural heritage of the Basque Country.

An important milestone in the history of the museum was its renovation and expansion. Since 1932, the exhibitions were held in the Dominican monastery of San Sebastián, an imposing building built in the 16th Century below the fortress of Urgull. The growing of the museum and the precarious state of the monastery made a renovation and the construction of a new building inevitable. The project was led by the Spanish architects Nieto and Sobejano, and thus brought a new design to the San Telmo Museum, which since its reopening in 2011 is called 'museum of the Basque society and citizenship' *(Museo de Sociedad Vasca y Ciudadanía)*.

The permanent exhibition covers five areas, which allow for a journey through the history of the Basque culture and society. The historical development of the building *(historia del sitio),* the challenges of society *(los desafios de nuestra sociedad),* the most important milestones in the history of the Basque Country *(huellas de la memoria),* the awakening of modernity *(despertar de la modernidad),* and the historical art collection *(collección histórica de arte)* are presented to the visitor.

As already indicated, the collections of the San Telmo Museum have grown since its inception, as did the demand of visitors from all over Spain, Europe and increasingly from other parts of the world such as China and Russia. Thus the museum has been challenged for several years to present its content as well and as accessibly as possible. Due to space and cost reasons, only two languages were chosen for the documentation and labeling in the exhibition rooms: Spanish and Basque, the two official languages of the Basque Country.

This initiative is appreciated by the locals and the educational institutions. On the one hand, the Basque language is considered an important part of Basque culture, on the other hand, especially since the introduction of Standard Basque in 1966, teaching in Basque is increasingly promoted. The material for the exhibitions is often translated in collaboration with other institutions, either initially written in Basque and then translated into Spanish, or vice versa.

The San Telmo Museum would like to appeal to visitors from other countries and win them over. Thus multilingualism plays an increasing role. Leaflets and audio guides are provided in French and English. On their website, at least the content of the main pages (the history of the museum, an overview of the exhibitions) is provided in French and English. For this, the museum cooperates not only with a company for multimedia equipment and facilities, such as the aforementioned audio guides or even short films, but also with a translation company that will be tasked with the

19 Some information is from the website www.santelmomuseoa.com and from the document "San Telmo MUSEUM.doc" which was provided by the museum staff.

translation of the texts from Spanish or Basque to French and English. The translations are manually created by freelance translators and checked by museum staff. The texts of the website are also updated and translated manually.

The museum had chosen English and French for the following reasons: English is the most widely spoken international language. French was chosen because of the high number of visitors from the neighboring country of France. The San Telmo Museum, however, is often criticized for having the texts in the exhibition rooms in Spanish and Basque only. Moreover, the demand for other languages, such as Italian, Russian and Chinese, is increasing. The Department of Communication is therefore trying to produce material (leaflets and audio guides) in additional languages for next summer. The main challenge is to find a cost effective and above all profitable form of production. For example, the addition of German is currently considered, but there are still doubts whether it would be worthwhile when considering the number of visitors.

In 2016, Donostia-San Sebastián will be the European Capital of Culture. At the latest for that occasion, the San Telmo Museum will expand its multilingualism. Currently the museum is in contact with a company in order to develop an application *(app)* for smartphones, which could eventually replace audio guides and is expected to attract an even wider audience.

Meanwhile, however, a major internal project is at hand, namely the temporary exhibition on the origins and the development of the Basque language in fall 2012.

6 EITB-Group in Bilbo-Bilbao (2012-06-07)

On thursday, we visited the EITB Broadcasting Company in Bilbao. EITB stands for *Euskal Irrati Telebista* and is the first Basque radio an television group. Today, its various information and entertainment products reach more than one million citizens everyday. The foundation of the EITB-Group was decided in 1982 by the Basque Parliament. Shortly thereafter, the first radio station went on the air and not much later television was launched as well. The EITB-Group is funded 80% by the Basque Government, the remaining revenue is generated through advertising revenue. The EITB-Group hat the promotion and dissemination of the Basque language as its target. This goes hand in hand with the standardization of the Basque language and the reform of public education. The use of the Basque language in public spaces was banned under Franco's dictatorship for a long time. In the more urban areas, such as Bilbao, the Basque language was suppressed almost completely by the Spanish. This led to a generation of Basques who could not speak Basque at all or only partially. In addition, Basque possesses a number of often very different dialects, which increased the barrier to its understanding. Besides the standardization and re-introduction of the Basque language in education and as official language, the media also helped its proliferation. This includes the governmentally funded EITB-Group. The problem with this sort of proliferation is that a whole generation is growing up who has only learned a standardized form of Basque and thus, still do not feel comfortable with the Basque which is spoken in the more rural regions. Standard Basque is perceived as something very artificial. The EITB-Group hopes to bridge that gap with their work and to support the Basque language further.

6.1 The channels of the EITB-Group

Today, the EITB-Group has five radio stations, five television channels and since 2008 a multilingual website (www.eitb.com) which available in Basque, Spanish, English and French. The TV channels *ETB-1*, entirely in Basque, and *ETB-2*, entirely in Spanish, have been on air since the 1980s. There are two international channels that are available via satellite and the internet: *ETB Sat* for Europe and *Canal Vasco* for America. Both transmit in Spanish and serve mainly to disseminate the Basque culture. However, they transmit sporting events as well. The news channel *ETB-3* is designed to appeal to a younger audience and sends content in Basque and Spanish. The addition of some segments in English is planned. Spain is a country which traditionally synchronizes all imported programs but in the scope of globalization the EITB-Group endeavors to promote the English language as well. The TV channels broadcast nonstop and there is an agreement with the government that a transmission fail is never allowed to happen.

The EITB-Group possesses a total of five radio stations, two generic stations, *Euskadi Irratia* in Basque and radio *Euskadi* in Spanish. *Gaztea* transmits in Basque and has a young target audience, whereas radio *Vitoria* serves as a local radio for the region Alava. Finally there is *EITB Musika* which plays music only.

6.2 The work at headquarters

The EITB-Group has locations in three cities: San Sebastián and Vitoria produce TV segments and

series; Bilbao is the headquarters with the editorial studios for radio, television and internet as well as other studios and production spaces. Newly occupied since 2007, the architecture of the headquarters is supposed to reflect the philosophy of the group with its open and airy design, and simultaneously, depict its work processes centralized and symbolically. First one enters the controlling area the heart of the television station. At this point of intersection, all incoming and outgoing content is bundled; this is where everything is monitored in order to prevent interruptions. Behind it is the brain, the computer where all content is processed, verified and stored. It is very important that the quality of the content is checked and that the correct metadata is added, in order for the journalists to search for content efficiently. All content is digitally archived. If something is not broadcasted within four days, it will be deleted forever. This serves to save storage space. If, however, the content is used, it is saved in a different location. The station is legally required to save everything they broadcasted for a time period of four years. In addition, the EITB-Group has a department concerned with the digitalization of old news pieces; this content finds its way into the archive as well.

Below and behind the control area is a large open room with workstations for journalists. The work stations include opportunities for research but also places for editing and sound editing. Originally, the newsroom had been planned as a combined editorial space for radio, TV and internet, however, it turned out that their tasks were too different. Therefore, the editorial offices are now separated and spatially divided. Nevertheless, is the open design supposed to help break barriers between employees and their superiors.

In addition, there are also several radio and television studios which produce live broadcasts. The life broadcasts are edited and subtitled live, because of a new law which requires subtitles for the hard of hearing. This of course leads to the generation of new data that is of interest for many different applications of computational linguistics and multilingual text analysis. Hence the subtitle production and the digitization of older reports can be seen as point of intersection for computational linguistics and they offer opportunities for further improvement of the production process.

7 Mixer Servicios Audiovisuales S.L. in Donostia-San Sebastián (2012-06-08)

Mixer Servicios Audiovisuales S.L. was founded in 2004, resulting from the cooperation between the two companies *Edertrack* and *Irusoin*[20]. Today, it is considered to be the most important dubbing and subtitle company in the Spanish-Basque region, consisting of two head offices in Bilbao and San Sebastián. Mixer employs 32 highly qualified technicians as well as more than 90 freelancer translators and actors for dubbing purposes. Furthermore, Mixer is a founder member of *EIKEN*, the Basque audiovisual group that constitutes a partnership of the most important Basque companies in the audiovisual field.

7.1 Services

The functions of the company are manifold. Besides the dubbing and subtitling of movies and other TV-programs, both of which are being more closely looked at and investigated further down, Mixer operates in the advertising industry, offers translations, recordings and soundtracks for teaching material, e-learning and advertisements, supports audio guides in the Basque language and partly also in Spanish for museums such as for instance the Guggenheim museum in Bilbao, and is operative in the post-production of sound and music in different movies and TV-programs. Since Mixer is a Basque company, its main focus in most duties lies on the Basque language.

7.1.1 Dubbing

A large part of the company's tasks deal with the dubbing of movies, TV-programs and cartoons, mostly for the Basque broadcasting service *Euskal Irrati Telebista* (EITB). The primary focus lies on dubbing from English to Basque or Spanish to Basque, but a smaller part also consists of dubbing from English into Spanish. For this purpose, Mixer offers an archive of different speakers that can be listened to on their website. Quite recently Mixer, together with Vicomtech, tried to transform the voice of a dubbing actor in such a way for it to be used for different parts in a movie. However, up to today this project could not be completed with satisfaction and is thus not in use.

The dubbing of films, DVD's, TV-programs and cartoons proves to be a very different task from the subtitling. Neither machine translation nor translation memory systems turn out to be helpful for the production of dubbing scripts, because on the one hand, it is necessary for the script to be in line with the mouth movements in the visual part, and on the other hand is it oftentimes difficult to transfer puns, jokes and metaphors from one language into the other not only with the meaning but also with the mouth movements in mind. As a rule, one translator, specially trained in dubbing, works for 1 to 2 weeks per film. For this task, the translator is given the film's transcript in the original language, as well as the movie itself in order to adjust the mouth movements. In the process, the translator appoints the exact time coding of the utterances and integrates these codes into the transcripts. Thereafter, the film is dubbed in the studio by actors on the basis of the dubbing script. The product is then reviewed by an editor before it is finally aired.

20 http://www.irusoin.com/

7.1.2 Subtitling

Besides the dubbing, subtitling is also a very important source of income for the Mixer company. For the most part, they deal with same-language subtitling of movies and TV-programs for the hard-of-hearing and deaf people and only a small part consists of translated subtitles. As in the field of dubbing, EITB is one of the larges customers for Mixer in this field as well, since approximately 75 percent of all broadcasted programs have to be provided with same-language subtitles this year and next year this number will increase to 90 percent of all programs. As opposed to dubbing, subtitles are generally semi-automatically generated with the help of software such as for instance *SuSa* (*semi-automatic subtitling*), which is a in-house development, established in cooperation with Vicomtech and which is used to convert the dubbing scripts into subtitles. For that purpose, the script and the audio file, including the time codes, are synchronized in *SuSa*, and the subtitler transfers the script into subtitles while watching the movie. The manually produced subtitles are then provided with the time code and formatted according to company regulations. In an additional step, further information for the hard-of-hearing as well as the deaf may be added such as for instance noises that are not explicitly part of the original transcript.

7.2 Conclusion

The visit to the company Mixer Servicios Audiovisuales S.L. in San Sebastián proved to be multifaceted and practice-oriented. Not only were the company and its products introduced, but they also provided us with a demonstration of the different programs used for the production of subtitles. In the second part of our visit, we were shown the studios in order to see how a TV-program is dubbed and produced.

Overall, there seems to be great potential for machine translation in this field. Since Spanish and English are syntactically very different from the Basque language, and since only sparse parallel data exists for Basque, it proves to be very difficult to automatically translate subtitles by means of a translation system, as it is successfully done in the project of the University of Zurich in collaboration with a Swedish subtitling company. Due to the fact that Mixer's main focus lies on same-language subtitles for the hard-of-hearing or deaf people, the task would have to be approached differently. For the future, it would be desirable for this industry - even if this probably will always remain an illusion - to establish a method to automatically adjust dubbing scripts to the mouth movements in the movie. Until that happens though, Mixer will certainly dub and translate many more movies.

8 Elhuyar in Donostia–San Sebastián (2012-06-08)

Elhuyar[21] was founded in 1972 with the goal to promote the use of the Basque language in science and technology. In 2002, Elhuyar was turned into a foundation, out of which two companies developed: *Elhuyar aholkuaritza* and *Eleka Ingeniaritza Linguistikoa*. The latter is a company that offers services in the field of speech technology. The name *Elhuyar* originates from the two Basque brothers José and Fausto Elhuyar, who jointly discovered the element Wolfram, the most important discovery of the Basque country.

8.1 Department for language services

The Elhuyar-Foundation is subdivided into different departments, one of it being the department for language services. The main goal for this department is to provide language resources, tools and services for the Basque as well as for other languages, such as for instance Spanish. The department for language services is in turn divided into three smaller departments: the department for translation and corrections, the department for the compilation of dictionaries, as well as the department for the research and development of language technologies and language resources. The focus of the research and the development lies in the field of terminology extraction, information extraction, corpora compilation, semantics, ontology, as well as machine translation. In the following, some of the main research areas are dealt with in more detail.

8.1.1 The compilation of corpora

Elhuyar's corpora are partly manually and partly automatically extracted from the Internet. One of the most important corpus of Elhuyar is the *ZT Corpusa*, a corpus of science and technology, which has been developed in collaboration with IXA. This corpus consists of Basque texts that have been published in the field of science and technology from 1990 to 2002. The corpus is composed of a balanced part (approx. 2M words) and of an unbalanced part (approx. 7M words). It is possible to either browse the whole corpus or simply the balanced part. The website[22] on which the corpus is accessible is frequented approximately 2000 times per month, mainly by lecturers or by editors of lexica.

Another Basque corpus, developed by Elhuyar, is the *Lexikoaren Behatokia*[23]. It is a monitor corpus, that is to say a corpus that is continuously complemented by new texts. The texts come from different media and are provided by the royal academy for the Basque language. The corpus is annually amended to up to 10 million words.

Furthermore, there is the corpus *CorpEus*[24], which uses the Internet as a corpus. The tool consists of a filter comprising Basque filter-words. Using these, all websites that correspond to the query and are in Basque, are extracted.

In addition, Elhuyar has complied a multilingual corpus, the so-called 'consumer-corpus'. It consists

21 http://www.elhuyar.org/EN/Elhuyar-Foundation
22 www.ztcorpusa.net
23 http://lexikoarenbehatokia.euskaltzaindia.net/cgi-bin/kontsulta.py
24 http://corpeus.elhuyar.org/cgi-bin/kontsulta.py

of texts taken from magazines of supermarkets (e.g. product comparisons) in the Basque, Catalan, Spanish and Galician languages. This corpus has been composed manually.

8.1.2 Terminology extraction

Elhuyar tries to automatically extract terminology from corpora in order to partially automatize the compilation of dictionaries. For this terminology extraction they use both linguistic and statistical methods and it can thus be described as a hybrid system. Since Basque is a highly agglutinating language and thus many different word forms exist, a purely statistical approach would be unsuitable.

In connection with the terminology extraction, the tool *Itzul Term*[25] has been developed, enabling the automatic extraction of corresponding Basque and Spanish words from Translation Memory systems. It is possible that this system will be extended to English as well. Since parallel corpora are currently only available for Basque-Spanish and Basque-English, they currently work with comparable corpora as well. Furthermore, a doctoral thesis is concerned with the extraction of (idiomatic) phrases from Basque corpora.

Another possibility to automatize the compilation of dictionaries is the Pivot-technology. If one would like to compile a dictionary for Basque-Chinese and it is known that dictionaries are available for the language pairs Basque-English and English-Chinese, English can be used as an intermediate language (so-called 'Pivot-language'). For a Basque word, all corresponding English words are collected, and then the corresponding Chinese words are assembled for these English words. These Chinese words now serve as translation possibilities for the Basque word. Finally, the corpora along with the context help assess which words are actually used.

8.1.3 Information extraction

Elhuyar developed *Elebila*[26], the first Basque search engine, as well as *Zientzia*[27]. It is a multilingual search engine: to begin with, Basque results are presented and then (also in case no Basque results are found) results from other languages (Spanish and English) are listed. The input is always in Basque, but during the search process, the input word is translated into English and Spanish in order to find the results in these languages as well. Problems of disambiguation try to be avoided by checking whether the documents are similar in terms of content.

Furthermore, research continues attempting to extract meaning from Basque texts, as well as automatically answered questions. A semantic web, in which information is saved, provides a basis for the answer to the question. The information is extracted from texts; in this case Elhuyar extracts the information from info boxes of the Basque DBpedia (a semantic web consisting of information from Wikipedia).

25 http://itzulterm.elhuyar.org
26http://www.elebila.eu/search_basic
27www.zientzia.net

8.1.4 Machine Translation

In the field of machine translation, Elhuyar works together with IXA. One product of this collaboration is the *OpenTrad-System*[28]. IXA developed the translation system and Elhuyar provided them with the essential data and corpora. Thus, Elhuyar did not develop a translation system on their own.

8.2 Conclusion

The research area of Elhuyar is very broad and they work with many companies and research centers that are concerned with the Basque language. The research of Elhuyar is application-oriented and deals primarily with Basque, but one of their goals is to offer their range of products in other languages as well (mainly Spanish and English). In this connection, the statement that Basque is much more complex than Spanish and English from a morphological point of view is very interesting. In other words, the tools that work for the Basque language, generally also work for Spanish or English.

28 http://www.opentrad.com//eu/inicio

9 A study week in the Basque Country – Conclusion

9.1 Basque – Opportunities and limitations of language technology

The Basque language faces a difficult situation concerning language technology for a variety of reasons.

Due to the language's isolated character, computational linguistics is not able to profit from a closeness to other languages and, therefore, cannot gain the same advantage as closely related language pairs when being translated from the source into the target language of the same language family.

Moreover, the highly agglutinative morphology of the Basque language carries fundamental difficulties, for instance the question about a word unit: What constitutes a word? In Basque a string usually consists of a variety of morphemes and thus frequently has to be aligned with several words of another, less agglutinative language. This creates a *One-To-Many-Matching.*

The difficulties are aggravated by the fact that Basque, which is composed of seven, in some cases very different, main dialects, has experienced efforts of standardization only since 1966; a process which was not completed until this day. For this reason language technology is not able to resort to a comprehensive literary heritage with a consistent linguistic usage which means that useful training data for projects such as machine translation is rare. For that purpose the *IXA-Group* draws on more recent texts such as text books, newspapers or technological literature. The amount of data as well as the selection of data are still greatly limited.

The training data problem is visible in the SUMAT project as well: Although the Basque corporation Vicomtech participates in the project, no subtitles are being translated into Basque – let alone translated from Basque. The reason for that is on the one hand the lack of availability of already existing Basque subtitles as training material for a machine translation system and, on the other hand, a necessity for a translation-language pair with Basque as target language does not exist. Since speakers of Basque are almost invariably raised in a multilingual environment, they are able to use Spanish subtitles instead.

Even though language technology in the Basque country contributes greatly to the promotion of the Basque language, its resources are greatly limited still. For instance, to date only a few Basque dictionaries and just one (open source) tagger exist.

Still, the visits to various companies in the field of language technology have shown that the efforts of Basque computational linguists have already yielded several remarkable projects.

Elhuyar Foundation (alongside many other things) developed a search engine which – supplemental to *Google* – is looking exclusively for Basque websites and texts on the internet. In *BerbaTek* Vicomtech has built an audiovisual Question-Answering-System for Basque and the *IXA-Group* with the team surrounding Gorka Labaka has built its own rule-based machine translation system (*Matxin*) for the language pair Spanish-Basque. These are just three examples of the many promising projects that were presented to us during this week.

Summing up it can be said that the standardization which the Basque language has been subjected to since 1966, has been a boon for technological language processing. The standardization is what made it possible in the first place to create normalized and uniform training data which are fundamental to many computational linguistics projects. More and more companies take on the language technological challenge dealing with the Basque language and, at the same time, encourage its accessibility by allowing it to be adapted to the modern means of communication.

9.2 How does language technology benefit from multilingualism? And how does multilingualism benefit from language technology?

In the past few years, the importance of Basque has increased greatly, it has since become an elective in schools and is asked for or even required in many positions. Moreover, the standardization of Basque (Euskara Batua) helped to promoted the use of Basque. Within the Basque Country, Basque was being encouraged and developed whereas at an international level it was of little significance (e.g. in comparison to Spanish). This is why practically no language technology resources or tools were developed for the Basque language. However, due to the positive development of Basque, the demand for language technology tools increased sharply in the Spanish Basque Country's local industry. For this reason and because of the strong identification of Basques with the Basque culture, the focus of many institutions we visited, now lies on the development of language technological resources for Basque. Most institutions also devote themselves to Spanish and some even to English.

Thus, on the one hand, language technology is promoted by multilingualism in the Basque Country by developing new applications for Basque and, on the other hand, by adapting existing language technology tools to Basque. Due to the uniqueness of Basque (see section 2.2) this presents the developers with new challenges. The solutions for Basque can be helpful for other languages and language technology issues, especially if the languages are structurally similar to Basque.

However, multilingualism benefits from language technology as well. The existing language technological resources facilitate the transfer between the languages Spanish and Basque (e.g. machine translation) and can assist in the acquisition of Basque as L2 (e.g. automatic correction systems or language learning systems). In addition, language technology tools such as the Basque search engine create an incentive to use Basque more frequently.

This week has shown us, that language technology and multilingualism influence each other positively. We learned that language technology concerns itself with languages with a low number of speakers. On the one hand, this creates a profit for language technology and on the other hand, it contributes to the preservation and promotion of minority languages.