Universität Zürich
MA Multilinguale Textanalyse

Bericht Intensivwoche MLTA
03.-09. Juni 2012
in Donostia-San Sebastián, Spanien

Wie die Sprachtechnologie von der Multilingualität profitieren kann – Wie die Multilingualität von der Sprachtechnologie profitieren kann: Eine Analyse am Beispiel des Baskenlandes

(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)

eingereicht am 24.06.2012

Inhaltsverzeichnis

1 Eine Intensivwoche im Baskenland - Einführung....4

2 Vicomtech in Donostia-San Sebastián (04.06.2012)....5

2.2 Das Baskische....5

2.3 Vicomtech....5

2.4 Human Speech and Language Technologies bei Vicomtech....6

2.5 Projekte bei Vicomtech....6

2.6 Fazit....8

3 Mondragon Lingua in Arrasate-Mondragón (04.06.2012)....9

3.1 Mondragón und die Mondragon Corporation Cooperativa....9

3.2 Mondragon Lingua....9

3.3 Mondragon Lingua und Maschinelle Übersetzung....10

4 Die IXA Research Group in Donostia-San Sebastián (05.06.2012)....11

4.1 Forschungsaktivität....11

4.2 KYOTO – Knowledge Yielding Ontologies for Transition-based Organization....12

4.3 OpeNER – Open Polarity Enhanced Named Entity Recognition....13

4.4 Maschinelle Übersetzung....13

4.5. Semantik-Analyse....14

4.6 Fazit....15

5 Das San Telmo Museum in Donostia-San Sebastián (06.06.2012)....16

6 Die EITB-Gruppe in Bilbo-Bilbao (07.06.2012)....18

6.1 Die Kanäle der EITB-Gruppe....18

6.2 Die Arbeit am Hauptsitz....19

7 Mixer Servicios Audiovisuales S.L. in Donostia-San Sebastián (08.06.2012)....20

7.1 Dienstleistungen....20

7.1.1 Synchronisation....20

7.1.2 Untertitelung....21

7.2 Fazit....22

8 Elhuyar in Donostia–San Sebastián (08.06.2012)....23

8.1 Abteilung für Sprachservices....23

8.1.1 Erstellung von Korpora....23

8.1.2 Terminologieextraktion....24

8.1.3 Informationsextraktion....25

8.1.4 Maschinelle Übersetzung....25

8.2 Fazit....25

9 Eine Intensivwoche im Baskenland - Fazit....26

9.1 Baskisch – Möglichkeiten und Grenzen der Sprachtechnologie....26

9.2 Wie profitiert Sprachtechnologie von Multilingualität? Und wie profitiert Multilingualität von Sprachtechnologie?....27

1 Eine Intensivwoche im Baskenland - Einführung

Die internationale Intensivwoche des spezialisierten Masterstudiengangs 'Multilinguale Textanalyse' (MLTA) beschäftigte sich mit der Frage, inwiefern die Sprachtechnologie von einer multilingualen Umgebung profitiert und welchen Nutzen die Multilingualität aus der aktuellen Sprachtechnologie ziehen kann. Da sich der Masterstudiengang im Allgemeinen mit mehrsprachigen Dokumenten und derer technologischer Bearbeitung sowie Verarbeitung beschäftigt, wurde die Untersuchung der komplexen Frage vom Zusammenhang der Technologie und der Sprachenvielfalt in der mehrsprachigen Region des Baskenlandes in Spanien durchgeführt. Das spanische Baskenland befindet sich an der Atlantikküste direkt an der Grenze zu Frankreich. Es setzt sich aus den Provinzen Gipuzkoa, Biskaya, Álava und Navarra zusammen. Im spanischen Baskenland gibt es zwei offizielle Amtssprachen. Neben dem Spanischen wird ebenfalls die regionale Amtssprache Baskisch verwendet. Es konnte keine Verwandtschaft zwischen dem Baskischen und einer anderen Sprache nachgewiesen werden. Nachdem die baskische Sprache zur Zeit der spanischen Diktatur weitestgehend unterdrückt wurde, erfährt sie seit 1978 im spanischen Teil des Baskenlands eine starke Stabilisierung in der Gesellschaft und besitzt heute in etwa 500.000 Sprecher in Spanien[1], welche sieben verschiedene baskische Dialekte sprechen. Die Standardisierung und Entwicklung einer baskischen Schriftsprache namens *Euskara Batua* führten dazu, dass Baskisch heute in der Bildung, Industrie, Verwaltung, Forschung und in den Medien des Baskenlandes stark vertreten ist.

Die an der Intensivwoche teilnehmenden acht Studenten besuchten vom 04. bis 08. Juni 2012 mit dem MLTA-Programmleiter Martin Volk und der Programm-Koordinatorin Jeannette Roth verschiedenste Institutionen und Unternehmen im Baskenland. Dieser Bericht gibt einen Einblick in die Besuche und die dort behandelten Themen und Projekte hinsichtlich der Frage der Beeinflussung von Multilingualität und Sprachtechnologie.

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

2 Vicomtech in Donostia-San Sebastián (04.06.2012)

2.2 Das Baskische

Am Anfang des Besuches der MLTA-Studenten bei Vicomtech wurde die Struktur der baskischen Sprache (baskisch *euskara*) von einer Linguistin vorgestellt. Es handelt sich um eine genetisch isolierte Sprache, die als prä-indogermanisch angesehen wird. Das Baskische besitzt eine stark agglutinierende Morphologie. Eine Charakteristik des Baskischen ist zudem die Polypersonalität. So können in einer Verbform mehrere Personen durch Affigierung markiert sein. Zudem besitzt das Baskische 16 Kasusmarkierungen, die durch Suffixe am Nomen realisiert werden. Dies führt zu einer sehr komplexen Verbmorphologie. Ein besonderes Merkmal des Baskischen ist die Ergativität. Enthält ein Satz ein intransitives Verb, so ist das Subjekt unmarkiert und steht im Absolutiv. Handelt es sich hingegen um einen transitiven Satz, so steht das Objekt im Absolutiv und ist unmarkiert, wohingegen das Subjekt im Ergativ steht und somit markiert ist.

2.3 Vicomtech

Vicomtech[2] ist ein Zentrum für angewandte Wissenschaft in den Bereichen der visuellen Interaktion und der Kommunikationstechnologie, das in Mitten des spanisch-baskischen Gebiets in der Stadt San Sebastián ansässig ist. Das Unternehmen entwickelt Technologien für Firmen und Institutionen und arbeitet somit eng mit Industriepartnern sowie anderen Institutionen aus der Region zusammen, wie beispielsweise IXA oder Elhuyar. Ausserdem ist Vicomtech an mehreren europäischen Projekten beteiligt, wie das weiter unten beschriebene SUMAT-Projekt. Das Unternehmen ist zudem Teil der Technologie-Allianz IK4, welche die 8 führenden baskischen Technologie-Zentren vereint. Mit ca. 100 nationalen und internationalen Mitarbeitern versucht Vicomtech in über 90 Projekten jährlich eine Brücke zwischen lokalen Unternehmen und der internationalen Wirtschaft zu schaffen. So sollen lokale baskische Firmen mit Vicomtechs Hilfe die Chance bekommen, von international aktueller technologischer Forschung zu profitieren.

2 www.vicomtech.org, Stand aller Online-Referenzen im Text: 23.06.2012

2.4 Human Speech and Language Technologies bei Vicomtech

Seit 2001 besitzt Vicomtech eine Abteilung für menschliche Sprache und Sprachtechnologie, welche von Dr. Arantza del Pozo Echezarreta geleitet wird, die Gastgeberin des Besuchs der MLTA-Studenten bei Vicomtech war. Die Abteilung entstand aus dem Bedarf an sprachtechnologischen Komponenten innerhalb der Forschung an audiovisuellen Anwendungen. Etwa zehn Wissenschaftler kümmern sich um die Entwicklung von Anwendungen zur natürlichen Sprachverarbeitung, die später von Unternehmen der Sprachindustrie im interdisziplinären Feld der Mensch-Maschine-Schnittstelle und des multilingualen Wissensmanagements genutzt werden. Dabei deckt die Abteilung drei Schwerpunkte ab: die Sprachverarbeitung (*Speech Processing*), die Verarbeitung natürlicher Sprache (*Natural Language Processing*) und den Bereich der Dialog-Systeme.

Im Rahmen des halbtägigen Besuchs der MLTA-Studenten bei Vicomtech wurden verschiedene Projekte durch die Wissenschaftler vorgestellt. Im Folgenden wird das abgeschlossene Projekt BerbaTek vorgestellt, die laufenden Projekte SUMAT und CAPER sowie das zukünftige Projekt OpeNER.

2.5 Projekte bei Vicomtech

BerbaTek

Das Forschungsprojekt BerbaTek[3], welches Ende 2011 abgeschlossen wurde, beschäftigte sich mit der Entwicklung von Sprach- und Multimediatechnologien. Das Ziel des Projektes war es, die Wettbewerbsfähigkeit von baskischen Firmen der Sprachindustrie zu verbessern. Dazu wurden vor allem Anwendungen für die Bereichen der Übersetzung, des Sprachenlernens und der Wissensverwaltung (c*ontent management*) erforscht. Das Team von Vicomtech entwickelte im Rahmen von BerbaTek beispielsweise ein audiovisuelles System für Question-Answering in Baskisch, das mithilfe eines Avatars dargestellt wird. Weiter wurde an der automatischen Synchronisation von Filmen gearbeitet. Ein offener Forschungspunkt blieb dabei die Stimmtransformation, welche zur Nutzung eines einzigen Synchronsprechers für verschiedene Rollen in einem Film beitragen sollte. Diese Anwendung konnte für die Industrie nicht zufriedenstellend entwickelt werden und bleibt ein offenes Forschungsfeld. Ausserdem erstellten Wissenschaftler ein paralleles Korpus für die Sprachen Baskisch und Spanisch, welches auf den zweisprachigen Gesetzes- und Verwaltungstexten des Baskenlandes basiert und unter anderem für die Entwicklung eines baskisch-spanischen maschinellen Übersetzungssystems genutzt wurde.

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

SUMAT

Das EU-Projekt SUMAT[4] arbeitet an einem Online-Service für Untertitel durch maschinelle Übersetzung (*An Online Service for Subtitling by Machine Translation*). Partner des Projekts ist neben Vicomtech und anderen europäischen Firmen auch das an der Universität Zürich entstandene Unternehmen TextShuttle. Der Übersetzungsservice für Untertitel wird für 9 verschiedene europäische Sprachen entwickelt. Am Ende soll der Benutzer zwischen 14 möglichen Sprachpaaren zur Übersetzung wählen können. Da immer mehr digitale Multimedia-Angebote wie z. B. Filme auf dem europäischen Markt angeboten werden, verlangt dies nach einer schnellen und qualitativ hochwertigen Möglichkeit, diese Angebote zur besseren Verständlichkeit mit Untertiteln zu versehen. Diese mehrsprachige Anforderung sieht das SUMAT-Projekt als eine Möglichkeit, die statistische maschinelle Übersetzung, zusammen mit humaner Nachbearbeitung durch Übersetzer, für die Untertitel-Industrie verfügbar zu machen. Der Online-Service soll zur Erstellung von Film-Untertitlen durch freiberufliche Übersetzer sowie Untertitel-Firmen effizient genutzt werden können und zur Steigerung der Produktivität dieser Industrie beitragen.

CAPER

Vicomtech ist neben SUMAT auch im EU-Projekt CAPER[5] involviert, welches die gemeinschaftliche Beschaffung, Verarbeitung, Erschliessung und Auswertung von Informationen zur Verhinderung von organisierter Kriminalität erforscht (c*ollaborative information acquisition, processing, exploitation and reporting for the prevention of organised crime*). Das Ziel des Projektes ist es, eine Plattform zu schaffen, auf der öffentliche und private Informationen über organisierte Kriminalität in Europa ausgetauscht werden können. Zur Beschaffung der Information werden multilinguale, audiovisuelle Inhalte und biometrische Informationen analysiert. Hierzu wird unter anderem die Technologie des Data-Minings genutzt. Auch greift das Projekt auf Techniken der Eigennamenerkennung (*named entity recognition*) zurück, um in Kriminalität involvierte Personen zu identifizieren und herauszufinden, wer in welche Aktionen involviert ist. Zur Verarbeitung dieser Namen erstellt das Projektteam eine NLP-Pipeline. Zusätzlich müssen die gesammelten Daten zum Austausch standardisiert werden. Das Projekt zielt darauf ab, die Endergebnisse benutzerfreundlich zu gestalten, sodass die Informationsplattform von Vollzugsbehörden (*Law Enforcement Agencies, LEAs*) in Europa genutzt werden kann. Eine Herausforderung für die Wissenschaftler ist es dabei, die multilingualen und multimodalen Inhalte mit Techniken der Semantikanalyse kompatibel zu machen und geeignete Tools für den Endbenutzer zu entwickeln.

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

OpeNER

Ab Juli 2012 werden Wissenschaftler von Vicomtech am EU-Projekt OpeNER[6] arbeiten. Das Ziel dieses Projekts ist es, durch Eigennamenerkennung und Sentiment-Analyse im Internet Meinungen über Orte aller Art der Tourismusbranche zu erkennen. Dabei sollen Bewertungen von Verbrauchern über Hotels, Städte, Restaurants, Sehenswürdigkeiten und andere touristische Orte gesammelt und den europäischen Tourismusagenturen sowie den betreffenden Unternehmen (Hotels, Restaurants etc.) bereitgestellt werden. Eine grosse Herausforderung dieses Projekts zeigt vor allem die multilinguale Untersuchung der gewählten Sprachen Spanisch, Englisch, Französisch, Deutsch, Holländisch und Italienisch. Als Partner des Projekts konnte das Online-Tourismus-Portal Tripadvisor gewonnen werden, welches wohl einen grossen Teil der zu analysierenden Orts-, Hotel- und Restaurant-Namen bereitstellen wird. Zudem sollen alle in diesem Projekt entwickelten Techniken und Werkzeuge am Ende als Open-Source-Tools bereitgestellt werden.

2.6 Fazit

Das baskische Unternehmen Vicomtech arbeitet in seiner Abteilung *Human Speech and Language Technologies* an vielen unterschiedlichen Forschungsprojekten und deckt damit viele Bereiche der aktuellen Sprachtechnologie ab. Doch nicht nur die Themenvielfalt, sondern auch die Zusammenarbeit mit vielen anderen Firmen, Institutionen und EU-Projekten und die damit zusammenhängenden Internationalisierung macht Vicomtech zu einem überzeugenden und interessanten Unternehmen. Eine positives Merkmal dieser Abteilung ist zudem die starke Integration von PhD-Studenten und Universitäts-Absolventen. Ausserdem spielt Vicomtech eine grosse Rolle in der Entwicklung von Sprachtechnologie für das Baskische. Anwendungen wie Untertitel-Generierung oder automatische Synchronisierung tragen zur Stabilisierung des Baskischen bei. Die Multilingualität des Baskenlandes und die Förderung der baskischen Sprache profitiert eindeutig von Vicomtechs Entwicklungen auf dem Gebiet der Sprachtechnologie.

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 (04.06.2012)

3.1 Mondragón und die *Mondragon Corporation Cooperativa*

Mondragón (baskisch *Arrasate*) ist eine kleine Stadt (knapp 20.000 Einwohner) in der baskischen Provinz Gipuzkoa und liegt ca. 70km weit von San Sebastián entfernt im Landesinnern.
Im Jahr 1943 legte der Priester José María Arizmendiarrieta das Fundament für die grosse Erfolgsgeschichte der Kleinstadt, indem er eine demokratisch organisierte Fachhochschule gründete. Nachdem diese Hochschule viele erfolgreiche Genossenschaften hervorgebracht hatte, wurde 1956 die *Mondragón Corporación Cooperativa* (MCC) gegründet, deren Hauptziel bis heute die systematische Schaffung von Arbeitsstellen ist. Die MCC zählt aktuell über 100 Unternehmen aus drei sehr unterschiedlichen Sektoren: dem Finanzsektor (z. B. Banken, Versicherungen), dem Industriesektor (z. B. Automobilindustrie) und dem Einzelhandel (z. B. Supermarktketten). 1997 errichtete die MCC sogar eine eigene, private Universität in Arrasate-Mondragón, die den baskischen Namen *Mondragon Unibersitatea* trägt. Diese zählt heute vier Fakultäten (Ingenieur-, Betriebs-, Gastronomie- und Geistes-/Erziehungswissenschaften) mit ca. 4000 Studenten. Ein grosser Vorteil der Universität liegt darin, dass die marktorientierten Studiengänge stark von der Nähe ihrer Institution zum Arbeitsmarkt profitieren können.

3.2 Mondragon Lingua

Im Dienstleistungszweig innerhalb des Industriesektors der MCC wurde 1973 das Sprachenzentrum CIM (*Centro de Idiomas Mondragón*) gegründet, das heute *Mondragon Lingua* heisst. Das Unternehmen hat sich seither einen Namen als Sprachschule gemacht, die schon anfangs der 80er-Jahre die Technologien der Informatik zur Unterstützung der Sprachkurse inkorporierte und deshalb als Pionier im Gebrauch von Computern für das Lehren von Sprachen gilt. Das Unternehmen zählt momentan 5 Zentren im Baskenland: in Bilbo-Bilbao, Donostia-San Sebastián, Arrasate-Mondragón, Gasteiz-Vitoria und Oñati-Oñate. Neben dem Sprachunterricht bietet Mondragon Lingua auch Übersetzungs- und Dolmetscherdienste an. Momentan arbeiten für Mondragon Lingua über 100 hochqualifizierte Übersetzer und Dolmetscher, die insgesamt in mehr als 40 Sprachen übersetzen.

3.3 Mondragon Lingua und Maschinelle Übersetzung

Aktuell verfolgt Mondragon Lingua ein Projekt, in welchem maschinelle Übersetzung für das Sprachpaar Englisch–Spanisch in den herkömmlichen Workflow der Human-Übersetzung integriert werden soll. Projektleiter sind Ane Ruiz de Zarate und Manuel de la Pascua. Die Trainingsdaten umfassen laut de la Pascua „mehrere 10 Millionen Wörter“ und als MT-System (*machine translation system*) wird auf Moses zurückgegriffen. Da das Projekt noch ganz am Anfang steht, können keine konkreten Aussagen über die genaue Vorgehensweise oder über den Erfolg des Projektes gemacht werden.

Während der Projektpräsentation durch die beiden Projektleiter wurde klar ersichtlich, dass die Zuwendung hin zur MT noch mit einem gewissen Widerwillen vonstatten geht (Zitat aus Powerpointpräsentation: „MT – because we have no other option“). Dennoch ziehen wir eine zentrale Erkenntnis aus dem Besuch bei Mondragon Lingua, nämlich dass die Übersetzungsindustrie immer mehr die Notwendigkeit der maschinellen Übersetzung erkennt und nach und nach auf den Zug aufsteigt.

4 Die IXA Research Group in Donostia-San Sebastián (05.06.2012)

Die Universität des Baskenlandes (spanisch *Universidad del País Vasco*, baskisch *Euskal Herriko Unibersitatea*) zählt zur Zeit etwa 45 000 Studenten und 5000 akademische Mitarbeiter.[7] Die rund 100 Abteilungen der Universität sind auf die vier Standorte Bilbao, Vitoria-Gasteiz, Eibar und San Sebastián verteilt. Die Universität ist zweisprachig: Etwa 35 % der Studenten studieren auf Baskisch.

Die in San Sebastián angesiedelte *IXA Research Group*[8] gehört zur Abteilung für Sprachen und Informatiksysteme. Sie wurde 1988 gegründet mit dem Ziel, grundlegende sprachtechnologische Ressourcen für das Baskische zu entwickeln und beschreibt ihr Forschungsgebiet heute etwas allgemeiner mit *Natural Language Processing*. Derzeit beschäftigt die Gruppe ein interdisziplinäres Team bestehend aus 33 Informatikern, 11 Linguisten und 3 technischen Assistenten. 28 promovierte Dozenten der Universität sind durch verschiedene Projekte mit IXA verbunden. 22 Dissertationen (davon 6 externe) werden hier betreut und 8 Studenten sind dabei abzuschliessen.

IXA arbeitet mit Organisationen wie Vicomtech oder der Elhuyar-Stiftung zusammen, sowie mit Universitäten in Barcelona, Amsterdam, Rom und Darmstadt.

4.1 Forschungsaktivität

In einem einführenden Vortrag gab uns Arantza Díaz de Ilarraza einen Überblick über die Aktivitäten von IXA. Die Forschungsgruppe beschäftigt sich mit allgemeinen Problemen der Sprachverarbeitung, besonders aber im Hinblick auf das Baskische. Sie ist bestrebt, die Sprachtechnologie für das Baskische strategisch zu fördern, indem linguistische (theoretische) Grundlagen und Ressourcen (z. B. Korpora) geschaffen werden, die die automatische Verarbeitung erst ermöglichen. Aber auch Werkzeuge für weitergehende Entwicklungen oder sogar fertige Anwendungen gehören zum Programm. Die Liste der hergestellten Produkte ist ansehnlich und umfasst Bereiche wie Information Retrieval und Informationsextraktion, Maschinelle Übersetzung, Sprachlern-Software, Morphologie, (Morpho-)Syntax und Lexikographie/Semantik.

Eines von IXAs Aushängeschildern ist der baskische Rechtschreibeprüfer *Xuxen* (bedeutet etwa ‘exakt, korrekt’), der für Textverarbeitungsprogramme (MS Office, OpenOffice) und die Mozilla-Produkte (Browser, E-Mail-Client) sowie als Online-Anwendung[9] kostenlos verfügbar ist. Zu betonen ist die Wichtigkeit eines Rechtschreibeprüfers für das Baskische: Viele Baskischsprecher sind sich in der Orthographie sehr unsicher, da es erst seit 1966[10] überhaupt eine standardisierte Schriftsprache gibt und die Sprache erst um 1980 Einzug in das Bildungssystem gefunden hat.

IXA war an der Erstellung verschiedener umfangreicher digitaler Ressourcen beteiligt. Für die Herstellung des Rechtschreibeprüfers wurde die EDBL (*Euskararen Datu-Base Lexikala*,

7 www.ehu.es
8 *ixa* (Aussprache etwa [iʃa]) ist der baskische Name für den Buchstaben X.
9 www.xuxen.com (nur Spanisch und Baskisch)
10 vgl. Handout der IXA Research Group vom 05.06.2012, Folie 9

‘Lexikalische Datenbank für das Baskische’) aufgebaut, die später auch für andere Anwendungen (Morphologieanalyse, Lemmatisierung) benutzt werden konnte. Sie umfasst rund 100 000 Lemmata mit morphologischer Information. Eine weitere lexikalische Ressource ist das baskische WordNet *BasWN*, also eine Wissensbasis, die Wortbedeutungen in strukturierte lexikalisch-semantische Beziehungen setzt. Es enthält knapp 60 000 Wörter mit insgesamt über 90 000 Bedeutungen (Synsets). Zwei baskische Korpora sind zudem entstanden: EPEC (*Euskararen Prozesamendurako Erreferentzia Corpusa*, ‘Referenzkorpus für die Verarbeitung des Baskischen’) ist ein 300 000 Wörter grosses Korpus, das manuelle Annotation bezüglich Morphologie, Phrasen, syntaktischer Dependenzen, semantischer Rollen und Synsets aus BasWN enthält. Das neuere, nur morphosyntaktisch annotierte Korpus ZTC (*Zientzia eta Teknologiaren Corpusa*, ‘Wissenschafts- und Technologie-Korpus’) ist mit 10 Millionen Wörtern deutlich grösser; von den automatisch erzeugten Annotationen wurde rund ein Fünftel von Hand korrigiert. Ein frei zugängliches Online-Interface[11] erlaubt komplexe Abfragen an das ZTC.

Die Grösse des Korpus erregte die Publikumsfrage, wie diese Menge von wissenschaftlichen Texten in baskischer Sprache verfügbar gemacht werden konnten, und ob überhaupt auf Baskisch publiziert werde. Technologische oder andere international interessante Publikationen sind denn auch mehrheitlich auf Englisch verfasst, solche von nationalem Interesse hingegen oft auf Spanisch. Es gibt aber auch Fachpublikationen, die ein baskischsprachiges Publikum erreichen. Die Universität, die einen beachtlichen Teil der Kurse parallel spanisch und baskisch anbietet, hat so auch einen hohen Bedarf an baskischer Fachliteratur. Digital war indes lange nicht genug Material verfügbar: Für das Korpus wurde viel Gedrucktes digitalisiert.

4.2 KYOTO – Knowledge Yielding Ontologies for Transition-based Organization

Aitor Soroa brachte uns das Projekt *KYOTO* näher, ein multilinguales Text-Mining-System im Bereich Umwelt- und Artenschutz. Das mit EU-Mitteln finanzierte Projekt hat eine offene, frei zugängliche Plattform für umweltbezogenes Wissen zum Ziel. Das Projekt wird von Prof. Piek Vossen der Freien Universität Amsterdam geleitet und schliesst Forschergruppen in Italien, Spanien, Tschechien, der Republik China, Japan und Deutschland ein.

Die Idee ist, dass das Text-Mining-System automatisch Ereignisse (Fakten) in Texten verschiedener Sprachen findet und sie so sprach- und kulturübergreifend zur Verfügung stellt.

Es unterstützt die Sprachen Spanisch, Englisch, Niederländisch, Italienisch, Chinesisch und Japanisch. Dabei fungiert WordNet als Sprachbrücke. Die nicht-englischen WordNets (vgl. EuroWordNet[12]) bauen strukturell auf dem englischen WordNet auf und verlinken auf dieses. So können die verschiedenen Ausdrücke auf die entsprechenden Konzepte verweisen.

In einem ersten Schritt erstellt das System eine Terminologie-Datenbank. Die sogenannten 'TyBots' *(Term Yielding Bots)* extrahieren Terme mit Hilfe von Wikipedia bilden sie auf Konzepte ab

11 www.ztcorpusa.net/cgi-bin/kontsulta.py (nur Baskisch)
12 www.illc.uva.nl/EuroWordNet/

und bauen eine Ontologie auf.
Die eigentliche Wissensextraktion erfolgt mit mehreren Hundert handgeschriebenen Regeln. Die sogenannten 'KyBots' *(Knowledge Yielding Bots)* generieren für jeden Satz ein Outputtemplate, wenn eine Regel matcht. Die Regeln enthalten jeweils sowohl eine morphosyntaktische Komponente als auch eine semantische Bedingung, die beide erfüllt werden müssen. Dies ist die schwierigste Aufgabe, da hier den vielfältigen Ausdrucksmöglichkeiten der natürlichen Sprache Rechnung getragen werden muss.
Das extrahierte Wissen liegt schliesslich auf drei Ebenen vor: lexikalisch (Terme), WordNet (Konzepte) und in der Ontologie. Die extrahierten Fakten werden in einem eigenen XML-Format (*KAF*, Kyoto Annotation Framework) gespeichert, das sich für linguistische wie semantische Weiterverarbeitung eignet und das flexibel und erweiterbar ist.

4.3 OpeNER – Open Polarity Enhanced Named Entity Recognition

Über das Projekt *OpeNER* berichtete uns Rodrigo Agerri. Das Projekt ist eine Zusammenarbeit mit Vicomtech und wird in Kapitel 2.5 näher beschrieben.

4.4 Maschinelle Übersetzung

Über die Aktivitäten in Maschineller Übersetzung in der IXA-Gruppe informierte uns Gorka Labaka. Das System *Matxin*, das einige von uns bereits kannten, wurde hier entwickelt. Es ist korpus- und regelbasiert und unterstützt die Sprachpaare Spanisch-Baskisch und Englisch-Baskisch.
Das neue hybride MT-System *SMatxinT* soll nun die Vorteile der statistischen MT in die Übersetzungen einbringen. Statistische MT zeichnet sich gegenüber regelbasierter MT in der Regel durch bessere lexikalische Abdeckung und Auswahl sowie besseren Umgang mit unbekannten Strukturen aus.
Im Vergleich zu statistischen MT-Systemen für die meisten westeuropäischen Sprachen bietet das Baskische als agglutinierende und damit äusserst formenreiche Sprache zusätzliche Schwierigkeiten. Dank der üppigen Flexion hat das Baskische eine sehr hohe Types/Tokens-Ratio, was zu Kargheit in den Trainingsdaten führt. Erschwerend kommt hinzu, dass die Verfügbarkeit von parallelen Daten für das Baskische begrenzt ist. Gleichzeitig kommt das Baskische in der Regel mit weniger Tokens pro Satz aus als beispielsweise die spanische Übersetzung, was zu häufigen 1:n-Alignierungen führt. Diese beeinträchtigen die Übersetzungsgenauigkeit tendenziell, da die Generalisierung so kleiner ausfällt.
Dieses Problem kann man entschärfen, indem ein Morphologiesystem die Wörter segmentiert und so in Stämme und Affixe oder Affixgruppen aufteilt. Durch kluge Wahl der Aufteilung können z. B. verbale Affixe im Baskischen mit Hilfsverben im Spanischen aligniert werden, was den Anteil an 1:1-Alignierungen erhöht und die Anzahl Types reduziert. So können auch Wortformen, die im Training nicht vorkamen, trotzdem korrekt übersetzt werden.

Das Hybridsystem SMatxinT verwendet ein regelbasiertes Subsystem als Gerüst für die Übersetzung. Mit Syntaxregeln wird die Struktur des Zielsatzes vorgegeben. Das statistische Subsystem errechnet dann für Phrasen oder andere Satzstücke mehrere alternative Übersetzungskandidaten zu denen des regelbasierten Systems, aus denen dann ausgewählt wird.

4.5. Semantik-Analyse

Der Nachmittag bot eine kurze Einführung in die wichtigsten und für die Projekte von IXA relevantesten Datensammlungen. Im Folgenden sollen einige der an diesem Nachmittag besprochenen Ressourcen kurz vorgestellt werden.

***General Inquirer* (1966)**

Das *General-Inquirer*-Projekt[13] wird auf seiner Homepage als computer-gestützter Ansatz für die Inhaltsanalyse von Texten und als Mapping-Tool beschrieben. Rodrigo Aguerri führte aus, dass es eine Ressource für Meinungen (*opinions*) sei, in welcher die beinhalteten Wörter zu positiven beziehungsweise negativen Werten kategorisiert sind. Spezifisch bedeutet dies, dass jede Textdatei nach Kategorien gemappt wird. Diese insgesamt 182 Kategorien bestehen aus einer Liste von Wörtern und Bedeutungen. Die Kategorie „negative“ ist nach Angabe des Projekts die grösste mit 2291 Einträgen. Nutzer haben die Möglichkeit, eigene beliebig grosse Kategorien zu erstellen. Das Tool ist jedoch monolingual, d.h. nur für Englisch verfügbar. Es ist eine der Ressourcen, welche in der Grundlagenforschung von IXA benutzt wurde.[14]

EuroWordNet

Das *EuroWordNet*[15] hatte das Ziel, eine mehrsprachige Datenbank mit WordNets für mehrere europäische Sprachen zu erstellen. Es wurde im Jahr 1999 fertiggestellt. Die verwendeten Sprachen sind Holländisch, Italienisch, Spanisch, Deutsch, Französisch, Tschechisch und Estnisch. Wie eine Vielzahl von WordNet-Projekten orientiert sich EuroWordNet am Princeton WordNet für amerikanisches Englisch. Die WordNets sind jeweils so aufgebaut, dass sie sich über eine Kategorie von Synonymen (Synsets) mit denen der anderen Sprachen verlinken lassen. Jede Sprache besitzt jedoch ihr spezifisches System zur Lexikalisierung. Die Verknüpfung der unterschiedlichen Sprachlexika geschieht durch den sogenannten Inter-Lingual-Index (durch Synonyme, semantische Beziehungen). Im EuroWordNet gibt es zusätzlich eine Top-Ontologie bestehend aus 63 semantischen Unterscheidungen, die allen Sprachen gemein ist. EuroWordNet wurde als Beispiel für ein grundlegendes WordNet vorgestellt, welches IXA als Inspiration und Ausgangslage für eigene Projekte, wie beispielsweise das baskische WordNet, diente.

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

SentiWordNet

Es existieren verschiedene Versionen von *SentiWordNet*[16] (1.0 und 3.0). Grundsätzlich ist es eine lexikalische Ressource für Opinion Mining. Das *SentiWordNet* schreibt jedem Synset eines WordNets drei Sentiment-Werte zu: Positivität, Negativität und Objektivität. Für die Forschung von IXA ist insbesondere das SentiWordNet 1.0 aus dem Jahr 2006 relevant.

WordNet-Affect

Das *WordNet-Affect*[17] ist eine Erweiterung eines anderen WordNet-Projekts: *WordNet Domains*. Wie der Name schon andeutet, enthält dieses WordNet eine zusätzliche Kategorie von Synsets, welche dazu dient, affektive Konzepte und Wörter darzustellen. Diese WordNet-Synsets erhielten affektive Bezeichnungen (*a-labels*), die auf gefühlsbezogene Konzepte angewendet wurden.

Q-WordNet

Das *Q-WordNet*[18] ist eines der (laufenden) Projekte von Rodrigo Agerri, welche sich mit Sentiment-Analyse befassen. Es ist auf seiner persönlichen Homepage als Download erhältlich. Dieses WordNet ist eine lexikalische Ressource, bestehend aus WordNet-Bedeutungen, welche automatisch mit positiver oder negativer Polarität klassifiziert wurden. Polaritätsklassifizierung wird vor allem eingesetzt um Meinungsforschung zu betreiben, wie beispielsweise im bereits vorgestellten *OpeNER*-Projekt. *Q-WordNet* wurde inspiriert von *SentiWordNet*. Im Unterschied zu *SentiWordNet* wurden die linguistischen Informationen nicht nur manuell, sondern zusätzlich auch noch von Hand annotiert. Für Rodrigo Agerri und sein Team brachte dies einen Mehraufwand, aber auch eine wichtige qualitative Verbesserung des Materials.

4.6 Fazit

Auf Mehrsprachigkeit wird bei der IXA-Forschungsgruppe grossen Wert gelegt. Als Universität des Baskenlandes ist Baskisch ein natürlicher Fokus von Ausbildung und Forschung. Das Baskische wird denn auch nicht nur in der Ausbildungstätigkeit, sondern auch durch Forschungsprojekte wie das Rechtschreibungsprüfungsprogramm *Xuxen* und die Erstellung digitaler Ressourcen (EDBL, BasWN, EPEC und ZTC) aktiv gefördert. Das maschinelle Übersetzungssystem *SMatxinT* leistet zudem über die beiden Sprachpaare Spanisch-Baskisch und Englisch-Baskisch einen Beitrag, Sprachgrenzen zu überbrücken. Laufende Projekte wie *KYOTO* und *OpeNER* erweitern die multilinguale Ausrichtung von IXA über westeuropäische Sprachen hinaus.

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

5 Das San Telmo Museum in Donostia-San Sebastián (06.06.2012)

Das San Telmo Museum in San Sebastián wurde im Jahre 1902 eröffnet und ist das älteste Museum im Baskenland. Zunächst umfasste es die Disziplinen der Kunst, Geschichte und Archäologie; eine wachsende Nachfrage und auch Notwendigkeit führte dazu, dass ein viertes Feld, die Ethnographie, hinzugefügt wurde.[19] Das Hauptanliegen des San Telmo besteht darin, das geschichtliche und soziokulturelle Erbe des Baskenlandes zu fördern.

Ein wichtiger Meilenstein in der Geschichte des Museums ist seine Renovierung und Erweiterung. Seit 1932 befanden sich die Ausstellungen im dominikanischen Kloster von San Sebastián, ein imposantes Gebäude, das im 16. Jahrhundert am Fusse der Festung Urgull errichtet wurde. Das Wachstum des Museums und der prekäre Zustand des Klosters führten zur Renovierung und Errichtung eines neuen Gebäudes. Das Projekt wurde von den spanischen Architekten Nieto und Sobejano geleitet, und somit auch zu einer neuen Gestaltung des San Telmo Museums, das sich selber seit der Neueröffnung in 2011 als 'Museum der baskischen Gesellschaft und Bürgerschaft' (*Museo de Sociedad Vasca y Ciudadanía*) bezeichnet.

Die permanente Ausstellung umfasst fünf Bereiche, die eine Reise durch die Geschichte der baskischen Kultur und Gesellschaft ermöglichen. Die historische Entwicklung des Gebäudes (*historia del sitio*), die Herausforderungen der Gesellschaft (*los desafios de nuestra sociedad*), die wichtigsten Meilensteine der Geschichte des Baskenlandes (*huellas en la memoria*), das Erwachen der Moderne (*despertar de la modernidad*), und die historische Kunstsammlung (*collección histórica de arte*) werden dem Besucher präsentiert.

Wie bereits angedeutet, hat der Umfang der Sammlungen des San Telmo Museums seit seiner Entstehung zugenommen. So auch die Nachfrage von Besuchern aus ganz Spanien, Europa und immer mehr auch aus anderen Teilen der Welt wie beispielsweise China und Russland. Das Museum steht somit schon seit mehreren Jahren vor der Herausforderung, seinen Inhalt so gut und zugänglich wie möglich zu gestalten. Aus Platz- und Kostengründen fiel die Entscheidung, welche Sprache für die Dokumentation und Beschriftung in den Ausstellungsräumen verwendet werden sollte, auf nur zwei Sprachen: Spanisch und Baskisch, die zwei offiziellen Sprachen des Baskenlandes.

Diese Initiative wird von Einheimischen und von den Erziehungsinstitutionen begrüsst. Einerseits wird die baskische Sprache als wichtiger Teil der baskischen Kultur angesehen; andererseits wird vor allem seit der Einführung der Standardform des Baskischen in 1966 das Unterrichten in Baskisch zunehmend gefördert. Das Material für die Ausstellungen wird oft in Zusammenarbeit mit anderen Institutionen entweder zunächst in Baskisch verfasst und dann intern ins Spanische übersetzt, oder umgekehrt.

Das San Telmo Museum möchte auch Besucher aus anderen Ländern ansprechen und für sich

19 Einige Informationen stammen von der Webseite www.santelmomuseoa.com und aus dem Dokument „SAN TELMO MUSEUM.doc“, das uns von den Mitarbeitern des Museums zur Verfügung gestellt wurde.

gewinnen. Somit spielt Multilingualität eine zunehmende Rolle. Flyers und Audioguides werden in Französisch und Englisch zur Verfügung gestellt. Auch auf der Internetseite wird zumindest der Inhalt der Hauptseiten (die Geschichte des Museum, den Überblick der Ausstellungen) in Französisch und Englisch angeboten. Hier besteht die Zusammenarbeit nicht nur mit einer Firma, die multimediale Geräte und Einrichtungen, wie die eben genannten Audioguides oder auch Kurzfilme zur Verfügung stellt, sondern auch mit einer Übersetzungsfirma, die damit beauftragt wird, die Texte vom Spanischen oder Baskischen ins Französische und Englische zu übersetzen. Die Übersetzungen werden manuell von freiberuflichen Übersetzern erstellt und von Mitarbeitern des Museums überprüft. Auch die Texte der Webseite werden manuell aktualisiert und übersetzt.

Die Wahl für Englisch und Französisch wird vom Museum wie folgt begründet: Englisch ist die meist gesprochene, internationale Sprache. Französisch wurde wegen der hohen Besucherzahlen aus dem Nachbarland Frankreich gewählt. Das San Telmo Museum wird allerdings oft dafür kritisiert, dass die Texte in den Ausstellungsräumen nur in Spanisch und Baskisch vorhanden sind. Ausserdem steigt die Nachfrage für andere Sprachen sowie Italienisch, Russisch und Chinesisch. Die Abteilung für Kommunikation ist daher zurzeit damit beschäftigt, Material (in Form von Flyern und Audioguides) in zusätzlichen Sprachen für den kommenden Sommer zu produzieren. Die Hauptherausforderung besteht darin, eine möglichst kostengünstige und vor allem rentable Form der Produktion anzuwenden. So wird beispielsweise im Moment in Betracht gezogen, auch Deutsch hinzuzufügen, jedoch bestehen noch Zweifel, ob dies in Anbetracht der Besucherzahlen schon lohnenswert wäre.

Im Jahr 2016 wird Donostia-San Sebastián die Kulturhauptstadt Europas sein. Spätestens für diesen Anlass wird das San Telmo Museum sein mehrsprachiges Angebot erweitern. Zurzeit steht das Museum in Kontakt mit einer Firma, um eine Anwendung (*App*) für Smartphones zu entwickeln, die eventuell die Audioguides ersetzen könnte und voraussichtlich ein noch grösseres Publikum ansprechen soll.

Vorerst steht aber auch noch ein wichtiges internes Projekt an, nämlich die temporäre Ausstellung über die Entstehung und Entwicklung der baskischen Sprache im kommenden Herbst 2012.

6 Die EITB-Gruppe in Bilbo-Bilbao (07.06.2012)

Am Donnerstag besuchten wir die EITB Broadcasting Company in Bilbao. EITB steht für *Euskal Irrati Telebista* und ist die erste baskische Radio- und Fernsehgruppe. Ihre verschiedenen Produkte zu Information und Unterhaltung erreichen heute täglich mehr als eine Million Bürger. Die Gründung der EITB wurde im Jahre 1982 vom baskischen Parlament beschlossen; kurz danach ging der erste Radiosender auf Sendung und nicht viel später startete der Fernsehsender. Die EITB-Gruppe wird zu 80 % von der baskischen Regierung finanziert, das restliche Einkommen wird durch Werbeeinnahmen generiert.

Die EITB-Gruppe hat die Förderung und Verbreitung der baskischen Sprache und Kultur zum Ziel. Dies geht Hand in Hand mit der Standardisierung des Baskischen und der Reform des Unterrichtswesen. Der Gebrauch der baskischen Sprache im öffentlichen Raum war unter Francos Diktatur lange Zeit verboten; in den urbaneren Gebieten, wie Bilbao, wurde das Baskische fast komplett vom Spanischen verdrängt. Dies führte zu einer Generation von Basken, die der Sprache nicht oder nur teilweise mächtig waren. Dazu kamen noch eine Reihe von teils sehr unterschiedlichen Dialekten, die die Verständigungsbarriere erhöhten. Neben der Standardisierung und Wiedereinführung des Baskischen als Amts- und Unterrichtssprache halfen auch die Medien bei der Wiederverbreitung des Baskischen. Dazu gehört auch die von der baskischen Regierung finanzierte EITB-Gruppe. Das Problem dabei ist, dass eine ganze Generation Basken heranwächst, die nur Standardbaskisch gelernt haben, aber sich immer noch nicht wohlfühlen mit der baskischen Sprache, während die Muttersprachler der ländlicheren Regionen das Standardbaskische als etwas sehr Künstliches empfinden. Die EITB-Gruppe hofft, mit ihrer Arbeit diese Lücke zu überbrücken und das Baskische weiter zu unterstützen.

6.1 Die Kanäle der EITB-Gruppe

Heute betreibt die EITB-Gruppe 5 Radiostationen, 5 Fernsehkanäle und seit 2008 eine multilinguale Website (www.eitb.com) verfügbar in den Sprachen Baskisch, Spanisch, Englisch und Französisch. Die Fernsehkanäle *ETB-1*, komplett in Baskisch, und *ETB-2*, komplett in Spanisch, sind seit den achtziger Jahren auf Sendung. Daneben gibt es zwei internationale Kanäle, die über Satellit und Internet erhältlich sind: *ETB Sat* für Europa und *Canal Vasco* für Amerika. Beide senden in Spanisch und dienen hauptsächlich der Verbreitung der baskischen Kultur, übertragen aber auch Sportevents. Der neuste Kanal *ETB-3* soll ein jüngeres Publikum ansprechen und sendet Inhalte in Baskisch und Spanisch, geplant sind aber auch einige Inhalte in Englisch. Spanien ist ein Land, das traditionellerweise alle importierten Programme synchronisiert, aber im Zuge der Globalisierung bemüht sich die EITB-Gruppe auch, die englische Sprache zu fördern. Die Fernsehkanäle senden durchgehend, und es besteht ein Abkommen mit der Regierung, dass niemals ein schwarzes Bild ausgestrahlt werden darf.

Die EITB-Gruppe besitzt insgesamt 5 Radiosender, zwei allgemeine Sender, *Euskadi Irratia* in

Baskisch und Radio *Euskadi* in Spanisch. *Gaztea* sendet in Baskisch und hat ein jugendliches Zielpublikum, während Radio *Vitoria* als Lokalradio für das Gebiet Alava dient. Zuletzt gibt es noch *EITB Musika*, wo nur Musik gespielt wird.

6.2 Die Arbeit am Hauptsitz

Die EITB-Gruppe besitzt Standorte in drei Städten: In San Sebastián und Vitoria werden Fernsehbeiträge und -serien produziert; in Bilbao befindet sich der Hauptsitz mit den Redaktionen von Radio, Fernsehen und Internet und weiteren Studios und Produktionsplätzen. Die Architektur des im Jahre 2007 neu bezogenen Hauptsitzes soll mit einem offenen und luftigen Design die Philosophie der Gruppe widerspiegeln; gleichzeitig sollen auch die Arbeitsprozesse zentralisiert und symbolisch dargestellt werden. Zuerst kommt man in den Kontrollbereich, dem Herzen des Fernsehsenders. In dieser Schnittstelle werden alle ein- und ausgehenden Inhalte gebündelt; hier wird kontrolliert, dass es zu keinen Unterbrüchen kommt. Dahinter befindet sich das Gehirn, die Computer, wo die Inhalte bearbeitet, verifiziert und gespeichert werden. Es ist sehr wichtig, dass die Qualität der Inhalte überprüft wird und dass korrekte Metadaten hinzugefügt werden, so dass ein Journalist die Inhalte besser finden kann. Alle Inhalte werden digital archiviert. Wenn ein Inhalt nicht innerhalb von vier Tagen tatsächlich zur Ausstrahlung kommt, wird der Inhalt für immer gelöscht. Dies dient zum Sparen von Speicherplatz. Wenn der Inhalt benutzt wird, wird er an einem anderen Ort gespeichert. Der Sender ist aus rechtlichen Gründen verpflichtet, alles, was tatsächlich ausgestrahlt wird, für 4 Jahre zu speichern. Ausserdem besitzt die EITB-Gruppe eine Abteilung zur Digitalisierung von älteren Nachrichten; diese Inhalte finden ebenfalls den Weg ins digitale Archiv.

Unter und hinter dem Kontrollbereich befindet sich ein grosser offener Raum mit Arbeitsplätzen für die Journalisten. Die Arbeitsplätze beinhalten Möglichkeiten zur Recherche, aber auch Plätze für Schnitt und Sound-Editing. Ursprünglich war ein Newsroom mit einer kombinierten Redaktion von Radio, Fernsehen und Internet geplant; aber es hat sich herausgestellt, dass die Arbeiten zu unterschiedlich sind. Daher sind die Redaktionen nun getrennt und räumlich aufgeteilt. Trotzdem soll das offene Design dabei helfen, Schranken zwischen Chefs und Angestellten abzubauen.

Zudem gibt es auch noch mehrere Radio- und Fernsehstudios, in welchen Livesendungen produziert werden. Die Livesendungen fürs Fernsehen werden live geschnitten und auch live untertitelt, da die neue Gesetzgebung Untertitel für hörbehinderte Menschen vorschreibt. Dies führt natürlich zur Generierung von neuen Daten, die für verschiedenste Anwendungen der Computerlinguistik und der multilingualen Textanalyse von Interesse sind. Daher sind die Untertitelproduktion sowie die Digitalisierung von älteren Berichten eine Schnittstelle zur Computerlinguistik und sie bieten Ansatzpunkte zur weiteren Verbesserung des Produktionsprozesses.

7 Mixer Servicios Audiovisuales S.L. in Donostia-San Sebastián (08.06.2012)

Mixer Servicios Audiovisuales S.L. wurde 2004 aus dem Zusammenschluss der Firmen *Edertrack* und *Irusoin* [20]gegründet und gilt heute als wichtigste Synchronisations- und Untertitelungsfirma in der spanisch-baskischen Region, mit zwei Firmensitzen in Bilbao und San Sebastián. Mixer beschäftigt 32 hoch qualifizierte Techniker, sowie über 90 freischaffende Übersetzer und Schauspieler für die Synchronisation. Mixer ist ebenfalls ein Gründungsmitglied von *EIKEN*, die baskische audiovisuelle Gruppe, welche eine Partnerschaft der wichtigsten baskischen Unternehmen aus dem audiovisuellen Bereich darstellt.

7.1 Dienstleistungen

Die Aufgabenbereiche der Firma sind sehr vielfältig. Neben der Synchronisation und Untertitelung von Filmen und anderen TV-Programmen, welche weiter unten genauer beschrieben werden, ist Mixer in der Werbebranche tätig, bietet Übersetzungen, Aufnahmen und musikalische Hinterlegungen für Unterrichtsmaterialien, e-Learning und Werbung an, unterhaltet Audioguides in Baskisch und zum Teil auch Spanisch für Museen wie zum Beispiel das Guggenheim Museum in Bilbao, und ist in der Nachbearbeitung von Ton und Musik in verschiedenen Filmen und TV-Programmen wirksam. Da es sich um ein baskisches Unternehmen handelt, liegt der Fokus in den meisten der Aufgabenbereichen auf der baskischen Sprache.

7.1.1 Synchronisation

Ein sehr grosser Teil der Aufgaben der Firma beschäftigen sich mit der Synchronisation von Filmen, TV-Programmen und Trickfilmen, vor allem für den baskischen Rundfunkdienst *Euskal Irrati Telebista* (EITB). Hauptsächlich wird vom Englischen ins Baskische oder vom Spanischen ins Baskische synchronisiert, ein kleinerer Teil besteht jedoch auch aus Synchronisationen vom Englischen ins Spanische. Mixer bietet dafür ein Archiv mit verschiedenen Sprechern an, die man sich auf ihrer Internetseite anhören kann. Vor einiger Zeit wurde in Zusammenarbeit mit Vicomtech versucht, durch Stimmtransformation die Stimme eines Synchronsprechers so zu verändern, dass ein Sprecher für verschiedene Rollen eingesetzt werden kann. Dieses Projekt konnte jedoch bis heute nicht zufriedenstellend entwickelt werden und ist zur Zeit nicht im Gebrauch.
Die Synchronisation von Kinofilmen, DVD's, TV-Programmen und Trickfilmen gestaltet sich sehr unterschiedlich von dem Aufgabenbereich der Untertitelung. Weder maschinelle Übersetzung noch Translation Memory Systeme erweisen sich bei der Herstellung von Synchronisationstranskripten als hilfreich. Grund dafür ist zum einen die Notwendigkeit, dass das Synchronisationstranskript jeweils auf die Mundbewegung im Visuellen abgeglichen ist, zum anderen ist es oftmals schwierig, Wortspiele, Witze und Metaphern von der einen in die andere Sprache zu transferieren und dabei zu bedenken, dass nicht nur der Sinn gleich bleiben muss, sondern auch die Mundbewegung und

20 http://www.irusoin.com/

das Bild übereinstimmen müssen. Pro Film arbeitet in der Regel ein auf Synchronisation spezialisierter Übersetzer für 1 bis 2 Wochen. Der Übersetzer bekommt dafür das Transkript des Filmes in der Originalsprache, sowie den Film, um die Mundbewegungen anpassen zu können. Dabei legt er die genauen Zeitkodierungen der Äusserungen fest und integriert diese im Transkript. Danach wird der Film im Studio von Schauspielern anhand des Synchronisationstranskripts synchronisiert und vom Editor überprüft, bevor es schliesslich ausgestrahlt wird.

7.1.2 Untertitelung

Neben der Synchronisation nimmt auch die Untertitelung einen grossen Aufgabenbereich ein bei der Firma Mixer. Der Hauptteil besteht dabei aus der gleichsprachigen Untertitelung von Filmen und TV-Programmen für schwerhörige oder taube Menschen und nur ein kleiner Teil besteht aus übersetzten Untertiteln. Wie bei der Synchronisation ist auch hier *EITB* einer der grössten Kunden von *Mixer,* da bei *EITB* dieses Jahr rund 75 Prozent aller ausgestrahlten Programme mit gleichsprachigen Untertiteln versehen werden müssen, im nächsten Jahr dann sogar rund 90 Prozent. Anders als bei der Synchronisation ist es hier gebräuchlich, die Untertitel semi-automatisch herzustellen mit Hilfe von Programmen wie zum Beispiel *SuSa* (*semi-automatic subtitling*), eine Eigenentwicklung, die in Zusammenarbeit mit Vicomtech hergestellt wurde und der Konversion des Synchronisationstranskripts in Untertitel dient. In *SuSa* wird dazu das Skript und das Audiofile mit dem Zeitkodierungen synchronisiert. Der Untertitler transferiert dann das Skript in Untertitel, während er sich den Film anschaut. Die eingefügten Untertitel werden dabei mit der Zeitkodierung des Filmes versehen und gemäss Richtlinien formatiert. In einem weiteren Schritt können danach noch weitere Informationen für schwerhörige und taube Menschen eingefügt werden, wie beispielsweise Geräusche, welche im ursprünglichen Transkript nicht explizit vorhanden sind.

7.2 Fazit

Der Besuch der Firma Mixer Servicios Audiovisuales S.L. in San Sebastián gestaltete sich als sehr vielfältig und praxisorientiert. Es wurden nicht nur die Firma und ihre Produkte vorgestellt, sondern auch verschiedene Programme zur Produktion von Untertiteln demonstriert. In einem zweiten Teil ermöglichte man uns, die Studios zu besichtigen, um ein Gefühl dafür zu bekommen, wie die Produktion einer synchronisierten TV-Sendung in der Praxis abläuft.

Allgemein scheint es noch grosses Potential für die maschinelle Übersetzung zu geben in diesem Bereich. Da die Sprachen Spanisch und Englisch syntaktisch sehr verschieden sind vom Baskischen, und zudem wenig parallel übersetztes Material zur Verfügung steht, ist es schwierig, die Untertitel mit Hilfe von einem maschinellen Übersetzungssystem automatisch zu übersetzen, wie es im Projekt der Universität Zürich zusammen mit einer schwedischen Untertitelfirma bereits gelungen ist. Da Mixer ausserdem hauptsächlich auf gleichsprachige Untertitel für schwerhörige oder taube Menschen fokussiert, müsste der Ansatz allgemein anders umgesetzt werden. Für die Zukunft wäre es natürlich wünschenswert für diese Industrie - auch wenn es wahrscheinlich eine Illusion bleiben wird -, dass es eine Möglichkeit gäbe, die Synchronisationsskripts automatisch auf die Mundbewegungen im Film abgleichen zu können. Bis das möglich ist, wird Mixer jedoch bestimmt noch einige Filme manuell transkribieren und übersetzen.

8 Elhuyar in Donostia–San Sebastián (08.06.2012)

Elhuyar[21] wurde 1972 gegründet, mit dem Ziel, die Verwendung der baskischen Sprache in der Wissenschaft und Technologie zu fördern. Im Jahr 2002 wurde Elhuyar dann zu einer Stiftung. Aus der Elhuyar Stiftung entstanden zwei Firmen: *Elhuyar aholkuaritza* und *Eleka Ingeniaritza Linguistikoa*. Die *Elhuyar aholkuaritza* will die Verwendung der baskischen Sprache fördern. Die *Eleka Ingeniaritza Linguistikoa* ist eine Firma, die Produkte und Dienstleistungen im Bereich der Sprachtechnologie anbietet. Der Name *Elhuyar* stammt von den zwei baskischen Brüdern José und Fausto Elhuyar, die gemeinsam das chemische Element Wolfram entdeckt haben. Dies stellt die wichtigste wissenschaftliche Entdeckung des Baskenlands dar.

8.1 Abteilung für Sprachservices

Die Elhuyar-Stiftung ist in verschiedene Abteilungen gegliedert, darunter gibt es auch eine Abteilung für Sprachservices. Das Hauptziel der Abteilung für Sprachservices besteht darin, Sprachressourcen, Tools und Services für das Baskische und andere Sprachen (z. B. Spanisch) zur Verfügung zu stellen. Die Abteilung für Sprachservices besteht wiederum aus Teildepartementen, eines für Übersetzungen und Korrektur, eines für die Erstellung von Wörterbüchern und eines zur Erforschung und Entwicklung von Sprachtechnologien und Sprachressourcen. Die Schwerpunkte für Forschung und für die Entwicklung von Sprachressourcen und Tools liegen in den Bereichen Terminologieextraktion, Informationsextraktion, Korporaerstellung, Semantik und Ontologien sowie maschineller Übersetzung. Im Folgenden wird auf einige dieser Forschungsschwerpunkte näher eingegangen.

8.1.1 Erstellung von Korpora

Die Korpora von Elhuyar werden teilweise manuell und teilweise automatisch aus dem Internet extrahiert. Ein wichtiges Korpus von Elhuyar ist das *ZT Corpusa*, das Wissenschafts- und Technologiekorpus, welches in Zusammenarbeit mit IXA entwickelt wurde. Dieses Korpus besteht aus baskischen Texten die von 1990 bis 2002 im wissenschaftlichen und technischen Bereich publiziert wurden. Das Korpus besteht aus einem balancierten Teil (ca. 2 Mio. Wörter) und einem unbalancierten Teil (ca. 7 Mio. Wörter). Es besteht die Möglichkeit entweder das ganze Korpus oder nur den balancierten Teil zu durchsuchen. Die Website[22] unter welcher das Korpus durchsucht werden kann, wird ungefähr 2000 Mal im Monat besucht, hauptsächlich von Dozenten oder von Verfassern von Lexika.

Ein weiteres baskisches Korpus von Elhuyar ist *Lexikoaren Behatokia*[23]. Dabei handelt es sich um ein Monitorkorpus, d.h. um ein Korpus, das ständig mit neuen Texten ergänzt wird. Die Texte stammen aus verschiedenen Medien und werden von der königlichen Akademie für die baskische

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

Sprache zur Verfügung gestellt. Das Korpus wird jährlich um bis zu 10 Mio. Wörter ergänzt.
Ausserdem gibt es das Korpus *CorpEus*[24]. Dabei wird das Internet als Korpus verwendet. Das Tool besteht aus einem Filter mit baskischen Filterwörtern. Mit diesem werden alle Internetseiten, die der Suchanfrage entsprechen und auf Baskisch verfasst sind, herausgefiltert.
Elhuyar hat auch ein multilinguales Korpus erstellt, das sogenannte 'Konsumentenkorpus'. Es besteht aus den Sprachen Baskisch, Katalanisch, Spanisch und Galicisch und setzt sich aus den Texten aus Supermarktmagazinen (z. B. Produktvergleiche) zusammen. Dieses Korpus wurde manuell zusammengestellt.

8.1.2 Terminologieextraktion

Elhuyar versucht, automatisch Terminologien aus Korpora zu extrahieren und so die Erstellung von Wörterbüchern teilweise zu automatisieren. Zur Terminologieextraktion werden sowohl linguistische als auch statistische Verfahren verwendet, es handelt sich also um ein hybrides System. Ein rein statistisches Verfahren wäre für das Baskische ungeeignet, da es sich um eine stark agglutinierende Sprache handelt, sodass sehr viele verschiedene Wortformen existieren.
Im Zusammenhang mit der Terminologieextraktion wurde das Tool *Itzul Term*[25] entwickelt, mit dem sich entsprechende baskische und spanische Wörter aus Translation Memory Systemen automatisch extrahieren lassen. Unter Umständen wird dieses System irgendwann für das Englische erweitert. Da nur parallele Korpora für Baskisch-Spanisch und Baskisch-Englisch vorhanden sind, wird momentan auch mit vergleichbaren Korpora experimentiert. Im Rahmen einer Doktorarbeit wird ausserdem versucht, (idiomatische) Phrasen aus baskischen Korpora zu extrahieren.
Eine andere Möglichkeit, um die Erstellung von Wörterbüchern zu automatisieren, ist die Pivot-Technik. Wenn beispielsweise ein Wörterbuch für Baskisch-Chinesisch erstellt werden soll und Wörterbücher für die Sprachpaare Baskisch-Englisch und Englisch-Chinesisch vorhanden sind, kann das Englische als Zwischensprache (sogenannte 'Pivot-Sprache') verwendet werden. Für ein baskisches Wort werden alle entsprechenden englischen Wörter gesucht, und für all diese englischen Wörter werden wiederum die entsprechenden chinesischen Wörter gesucht. Diese kommen als Übersetzungskandidaten in Frage. Welche Übersetzungskandidaten dann tatsächlich übernommen werden, wird mit Korpora unter Einbezug des Kontextes geprüft.

24 http://corpeus.elhuyar.org/cgi-bin/kontsulta.py
25 http://itzulterm.elhuyar.org

8.1.3 Informationsextraktion

Elhuyar hat *Elebila*[26], die erste baskische Suchmaschine, entwickelt, sowie *Zientzia*[27]. Dabei handelt es sich um eine mehrsprachige Suchmaschine: Zuerst werden baskische Resultate angezeigt und danach (oder falls keine baskischen Resultate gefunden werden), werden die Resultate anderer sprachen (Spanisch und Englisch) aufgelistet. Der Input ist immer Baskisch, aber beim Suchprozess wird das eingegebene Wort ins Englische und Spanische übersetzt, um auch die Resultate in diesen Sprachen zu finden. Probleme bei der Disambiguierung versuchen sie zu vermeiden, indem geprüft wird, ob die Dokumente vom Inhalt her ähnlich sind.
Ausserdem wird daran geforscht Meinungen aus baskischen Texten zu extrahieren und Fragen automatisch zu beantworten. Als Basis für die Beantwortung dieser Fragen dient ein Semantisches Netz, in welchem die Informationen gespeichert werden. Die Informationen werden aus Texten extrahiert, Elhuyar extrahiert die Informationen aus den Infoboxen des baskischen DBpedia (= Semantic Web der Informationen aus Wikipedia).

8.1.4 Maschinelle Übersetzung

Im Bereich der maschinellen Übersetzung arbeitet Elhuyar mit IXA zusammen, daraus entstand das OpenTrad-System[28]. Die IXA entwickelte das Übersetzungssystem und Elhuyar lieferte die notwendigen Daten und Korpora. Elhuyar hat also kein eigenes Übersetzungssystem entwickelt.

8.2 Fazit

Elhuyar forscht in sehr vielen Bereichen und arbeitet mit vielen Firmen und Forschungszentren, die sich mit dem Baskischen beschäftigen, zusammen. Die Forschung von Elhuyar ist anwendungsorientiert und beschäftigt sich primär mit dem Baskischen, das Angebot soll aber auch für andere Sprachen ausgebaut werden (v.a. Spanisch und Englisch). In diesem Zusammenhang war die Aussage interessant, dass Baskisch morphologisch gesehen viel komplexer ist als das Spanische und Englische. Mit anderen Worten: Diejenigen Tools, die fürs Baskische funktionieren, funktionieren dann in der Regel auch für das Spanische oder Englische.

26 http://www.elebila.eu/search_basic
27 www.zientzia.net
28 http://www.opentrad.com//eu/inicio

9 Eine Intensivwoche im Baskenland - Fazit

9.1 Baskisch – Möglichkeiten und Grenzen der Sprachtechnologie

Das Baskische hat in der Sprachtechnologie aus verschiedenen Gründen einen schwierigen Stand.

Aufgrund des isolierten Charakters der Sprache kann die Computerlinguistik zum Beispiel nicht von einer Nähe zu anderen Sprachen profitieren und somit auch keine Vorteile aus eng verwandten Sprachpaaren ziehen, wie das bei Übersetzungen aus einer Quell- in eine Zielsprache aus derselben Sprachfamilie der Fall ist.

Die stark agglutinierende Morphologie der baskischen Sprache bringt ebenfalls grundlegende Probleme mit sich; so zum Beispiel die Frage nach der Worteinheit: Was ist ein Wort? Ein String besteht im Baskischen meist aus einer Vielzahl an Morphemen und muss demnach häufig mit mehreren Worten einer anderen, weniger stark agglutinierenden Sprache aligniert werden, es entsteht also ein *One-To-Many-Matching*.

Erschwerend kommt hinzu, dass das Baskische, welches sich aus sieben teilweise sehr unterschiedlichen Hauptdialekten zusammensetzt, erst seit 1966 eine Standardisierung erfahren hat, wobei dieser Vorgang bis heute nicht abgeschlossen ist. Aus diesem Grund kann die Sprachtechnologie auch nicht auf ein umfassendes Literaturerbe mit einheitlicher Sprachnorm zurückgreifen, weshalb die Trainingsdaten für Projekte wie Maschinelle Übersetzung wenig umfangreich sind. Für solche Zwecke greift beispielsweise die *IXA-Group* hauptsächlich auf jüngere Texte wie Sprachbücher, Zeitungen oder technologische Fachliteratur zurück. Die Datenmengen sowie die Auswahl der Daten sind jedoch stark eingeschränkt.

Das Problem der Trainingsdaten zeigt sich auch im SUMAT-Projekt: Obwohl Vicomtech als baskisches Unternehmen an dem Projekt beteiligt ist, werden keine Filmuntertitel ins Baskische – geschweige denn aus dem Baskischen – übersetzt. Der Grund dafür ist einerseits die fehlende Verfügbarkeit von bereits bestehenden baskischen Untertiteln als Trainingsmaterial für das Übersetzungssystem, andererseits besteht erst gar nicht die Notwendigkeit eines Übersetzungs-Sprachpaares mit Baskisch als Zielsprache. Da die Sprecher des Baskischen fast ausnahmslos mehrsprachig aufwachsen, können sie sich somit der spanischen Untertitel bedienen.

Obwohl die Sprachtechnologie im Baskenland einen grossen Aufwand zur Förderung der baskischen Sprache leistet, sind die sprachtechnologischen Ressourcen noch stark limitiert. So existieren bis heute nur wenige baskische Wörterbücher und sogar nur ein einziger (Open-Source-) Tagger.

Die Besuche bei den verschiedenen baskischen Unternehmen im Bereich der Sprachtechnologie haben aber gezeigt, dass die Bemühungen der baskischen Computerlinguisten bereits viele bemerkenswerte Projekte hervorgebracht haben. So hat die *Elhuyar Foundation* (nebst vielem anderem) eine Suchmaschine entwickelt, die – ergänzend zu *Google* – ausschliesslich nach

baskischen Websites und Texten im Internet sucht. Mit *BerbaTek* hat Vicomtech u.a. ein audiovisuelles Question-Answering-System fürs Baskische gebaut und die *IXA-Group* hat mit einem Team rund um Gorka Labaka ein eigenes, regelbasiertes maschinelles Übersetzungssystem (*Matxin*) für das Sprachpaar Spanisch–Baskisch gebaut. Dies sind nur drei Beispiele aus zahlreichen vielversprechenden Projekten, die uns in dieser Woche vorgestellt wurden.

Zusammenfassend lässt sich festhalten, dass die Standardisierung, welcher die baskische Sprache seit 1966 unterzogen wird, einen Segen für die technologische Sprachverarbeitung darstellt. Sie erlaubt es erst, dass nach und nach normalisierte, einheitliche Trainingsdaten entstehen, welche für viele computerlinguistische Projekte grundlegend sind. So nehmen sich auch immer mehr Unternehmen der Herausforderungen im sprachtechnologischen Umgang mit der baskischen Sprache an und fördern dieselbe gleichzeitig, indem sie einen Zugang zur Sprache ermöglichen, der den modernen Kommunikationsmitteln angepasst ist.

9.2 Wie profitiert Sprachtechnologie von Multilingualität? Und wie profitiert Multilingualität von Sprachtechnologie?

In den letzten Jahren hat die Wichtigkeit des Baskischen stark zugenommen, da es in der Schule zum Pflichtfach wurde und für viele Anstellungen mittlerweile erwünscht oder sogar vorausgesetzt wird. Auch die Standardisierung des Baskischen (Euskara Batua) trug dazu bei, die Verwendung des Baskischen zu fördern. Innerhalb des Baskenlands wird das Baskische gefördert und ausgebaut, während es auf internationaler Ebene (z. B. im Vergleich zum Spanischen) kaum Bedeutung findet. Deswegen wurden in anderen Gebieten Spaniens und anderen Ländern praktisch keine sprachtechnologischen Ressourcen oder Tools für das Baskische entwickelt. Durch die positive Entwicklung des Baskischen stieg die Nachfrage nach sprachtechnologischen Tools für das Baskische in der regionalen baskischen Industrie jedoch stark an. Aus diesem Grund und aufgrund der starken Identifikation der Basken mit der baskischen Kultur, liegt der Schwerpunkt vieler Institutionen, die wir besuchten, auf der Entwicklung von sprachtechnologischen Ressourcen für das Baskische. Die meisten Institutionen beschäftigen sich daneben noch mit dem Spanischen und teilweise mit dem Englischen.

Die Sprachtechnologie wird also durch die Multilingualität im Baskenland gefördert, indem einerseits neue Anwendungen für das Baskische entwickelt werden und andererseits existierende sprachtechnologische Tools für das Baskische angepasst werden. Aufgrund der Einzigartigkeit des Baskischen (siehe Kapitel 2.2) stellt dies die Entwickler vor neue Herausforderungen. Die Lösungsansätze für das Baskische können jedoch auch für andere Sprachen und sprachtechnologische Fragestellungen hilfreich sein, vor allem, wenn es sich um dem Baskischen strukturell ähnliche Sprachen handelt.

Die Multilingualität profitiert jedoch auch von der Sprachtechnologie. Die gegebenen

sprachtechnologischen Ressourcen erleichtern den Transfer zwischen den beiden Sprachen Spanisch und Baskisch (z. B. maschinelle Übersetzung) und können beim Erwerb des Baskischen als L2 helfen (z. B. automatische Korrektursysteme oder Sprachlernsysteme). Ausserdem schaffen sprachtechnologische Tools, wie die baskische Suchmaschine, einen Anreiz, das Baskische häufiger zu verwenden.

Die Woche hat uns gezeigt, dass sich Sprachtechnologie und Multilingualität gegenseitig positiv beeinflussen. Wir haben gesehen, dass sich die Sprachtechnologie auch mit Sprachen mit geringer Sprecheranzahl beschäftigt. Dadurch entsteht einerseits ein Profit für die Sprachtechnologie und andererseits kann ein Beitrag zum Erhalt und zur Förderung von Minderheitensprachen geleistet werden.