Université de Zurich
MA Analyse Multilingue de Texte

Rapport de la semaine intensive MLTA
03-09 juin 2012
à Donostia-San Sebastián, Espagne

Le bénéfice de la technologie du langage pour le plurilinguisme - Le bénéfice du plurilinguisme pour la technologie du langage L'exemple du Pays basque

(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)

Version française
16 septembre 2012

1 Une semaine intensive au pays basque – Introduction

La semaine intensive internationale du master spécialisé 'Analyse multilingue de texte' (MLTA) abordait les questions suivantes : Comment la technologie du langage peut-elle profiter d'un environnement multilingue ? Le plurilinguisme peut-il bénéficier des actuelles technologies du langage ?

Le master se penche principalement sur les documents multilingues et sur leur élaboration et traitement technologique. C'est pourquoi, l'intérêt principal portait sur les relations complexes qui existent entre la technologie et la diversité des langues présentes dans la région plurilingue du Pays basque en Espagne. Le Pays basque en Espagne se situe sur la côte atlantique avoisinante à la France. Il est composé des provinces Gipuzkoa, Biskaya, Álava et Navarra. Il y a deux langues officielles: l'espagnol et la langue officielle régionale basque. Il n'existe pas de parenté attestée entre le basque et une autre langue. Cette première a été réprimée pendant les années de la dictature espagnole ; cependant, depuis 1978, la langue basque s'est fortement établie dans la société et elle compte aujourd'hui environ 500.00 locuteurs en Espagne[1], qui parlent sept dialectes différents du basque. La standardisation et le développement d'une langue écrite nommée *Euskara Batua* ont résulté dans une forte présence du basque dans l'éducation, l'industrie, l'administration, la recherche scientifique et dans les médias.

Les huit étudiants participant à la semaine intensive, accompagnés du directeur MLTA Martin Volk et de la coordinatrice MLTA Jeannette Roth, ont visité plusieurs institutions et entreprises dans le Pays basque entre le 4 et le 8 Juin 2012. Dans ce rapport, les visites ainsi que les sujets et projets abordés sont décrits en tenant compte de la relation entre le plurilinguisme et la technologie du langage.

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

2 Vicomtech à Donostia – San Sebastian (04.06.2012)

2.1 La langue basque

Au début de la visite des étudiants MLTA chez Vicomtech, une linguiste a donné une présentation sur la structure de la langue basque (nommée *euskara* en basque). La langue est génétiquement isolée et est considérée comme pré-indo-européenne. Le basque se caractérise par une forte morphologie agglutinante. Un autre trait de la langue basque est sa polypersonnalité. Un verbe peut contenir plusieurs affixes qui représentent plusieurs personnes. De plus, le basque a 16 marquages de cas, réalisés par des suffixes nominaux. Cela résulte en une morphologie complexe du verbe. Une caractéristique spéciale de la langue basque est l'ergativité : si une phrase contient un verbe intransitif, le cas du sujet non marqué est l'absolutif ; par contre, s'il s'agit d'une phrase transitive, c'est l'objet non marqué qui porte le cas de l'absolutif alors que le cas du sujet marqué est l'ergatif.

2.2 Vicomtech

Vicomtech[2] est un centre pour les sciences appliquées dans les domaines de l'interaction visuelle et de la technologie de la communication. Le centre est situé au milieu de la région basque en Espagne dans la ville de San Sebastián. L'entreprise développe des technologies pour d'autres entreprises et institutions. Le travail de Vicomtech se fait donc en collaboration étroite avec des partenaires industriels et d'autres institutions de la région basque, comme par exemple IXA ou Elhuyar. Par ailleurs, Vicomtech participe à plusieurs projets européens, comme le projet SUMAT, qui sera décrit plus bas. En outre, l'entreprise fait partie de l'alliance IK4, qui unit les 8 centres technologiques basques de premier plan. Avec ses 100 employés nationaux et internationaux , et ses 90 projets annuels, Vicomtech essaye de créer un lien entre les entreprises locales et l'économie internationale. Ainsi, les entreprises basques ont l'opportunité, avec l'aide de Vicomtech, de profiter de la recherche technologique actuelle internationale.

2 www.vicomtech.org, toutes les références en ligne ont été enregistrées le 11.08.2012

2.3 Human Speech and Language Technologies chez Vicomtech

Depuis 2001, Vicomtech dispose d'un département pour le langage naturel et la technologie du langage, dirigé par Dr. Arantza del Pozo Echezarreta, qui fut également l'hôte de la visite des étudiants MLTA chez Vicomtech. La création du département provient du besoin croissant en technologies du langage dans le domaine de la recherche des applications audiovisuelles. Environ 10 scientifiques sont chargés du développement des applications pour le traitement naturel du langage. Ces applications sont ensuite utilisées dans les entreprises de l'industrie du langage, dans le cadre interdisciplinaire de l'interaction homme-machine et dans le cadre du management multilingue du savoir. Le département couvre trois thèmes principaux: le traitement du langage (*Speech Processing*), le traitement des langues naturelles (*Natural Language Processing*), ainsi que le domaine des systèmes de dialogue.

Pendant la visite des étudiants MLTA chez Vicomtech, plusieurs présentations des projets ont été données par les chercheurs de Vicomtech. Dans la suite, le projet fini BerbaTek sera décrit, tout comme les projets courants SUMAT et CAPER et le projet futur OpeNER.

2.4 Les projets chez Vicomtech

BerbaTek

Le projet de recherche BerbaTek[3], qui a pris fin en 2011, traite du développement de la technologie du langage et de la technologie multimédia. Le but de ce projet était d'améliorer la compétitivité des entreprises spécialisées en technologie du langage au Pays basque. La recherche s'est surtout penchée sur les systèmes de traduction, d'apprentissage des langues et du *content management*. Dans le cadre du projet BerbaTek, le groupe des chercheurs de Vicomtech a développé par exemple un système/logiciel audiovisuel pour le *Question-Answering* en basque, qui est représenté à l'aide d'un avatar. En outre, les chercheurs ont travaillé sur la synchronisation automatique de films. La transformation de voix, qui avait pour but l'utilisation d'une seule personne pour plusieurs rôles dans un film, est restée un point de recherche inaccompli dans ce projet. Le développement de ce système n'a pas pu satisfaire les partenaires industriels. Cependant, l'équipe de BerbaTek a établi un corpus parallèle pour le basque et l'espagnol, qui est formé à partir des textes administratifs bilingues du Pays basque. Ce corpus a été utilisé pour l'élaboration d'un système de traduction automatique entre le basque et l'espagnol.

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

SUMAT

Le projet européen actuel SUMAT[4] développe un service en ligne pour la traduction automatique de sous-titres (An Online Service for Subtitling by Machine Translation). Le projet a plusieurs partenaires : l'entreprise Vicomtech et d'autres firmes européennes, mais aussi l'entreprise TextShuttle, qui a été fondée à l'Université de Zurich. Le service de traduction de sous-titres compte 9 langues européennes. À la fin du projet, l'utilisateur aura le choix entre 14 paires de langues pour la traduction. Etant donné que de plus en plus de produits multimédia, comme par exemple des films, apparaissent sur le marché européen, une demande croissante de méthodes rapides et qualitatives pour permettre une meilleure compréhension à l'aide de sous-titres est recensée partout en Europe. Pour le projet SUMAT, cette demande en plurilinguisme signifie de pouvoir mettre à disposition de l'industrie la traduction automatique du sous-titrage, en combinaison avec la postproduction humaine par des traducteurs. Le service en ligne sera donc principalement utilisé pour la production de sous-titres par des traducteurs, mais il vise aussi à augmenter la productivité de l'industrie du sous-titrage.

CAPER

A part SUMAT, l'entreprise Vicomtech participe aussi au projet européen CAPER[5], qui explore l'acquisition, le traitement, l'exploitation et l'interprétation collaborative des informations pour la prévention de la criminalité organisée (*collaborative information acquisition, processing, exploitation and reporting for the prevention of organised crime*). Le but du projet est de créer une plateforme sur laquelle il sera possible d'échanger des informations publiques et privées sur la criminalité organisée en Europe. Pour l'approvisionnement des informations, des contenus multilingues et audiovisuels seront analysés avec des informations biométriques. A cet égard, la technologie du *data mining* sera utilisée. Le projet recourt aussi aux techniques de l'identification des noms propres (*named entity recognition*), pour identifier les personnes qui sont impliquées dans la criminalité et pour établir en quelles actions elles sont impliquées. Pour le traitement des ces noms propres, l'équipe du projet élabore un pipeline NLP. De plus, les données collectées devront être standardisées pour l'échange. Le projet vise à présenter les résultats de façon conviviale, pour que la plateforme des informations puisse être utilisée par les agences d'exécution en Europe. C'est un grand défi pour les chercheurs de combiner les contenus multilingues et multimodaux avec les techniques de l'analyse sémantique et de développer des outils appropriés pour les utilisateurs.

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

OpeNER

A partir de Juillet 2012, des chercheurs de Vicomtech vont travailler sur le projet OpeNER[6]. Le but de ce projet est de découvrir les avis en ligne sur toute sorte d'endroits touristiques, à l'aide de l'identification des noms propres et de la *sentiment analysis*. L'équipe veut rassembler les évaluations d'hôtels, de villes, restaurants, attractions touristiques et d'autres sites touristiques, faites auparavant par des consommateurs. Ces évaluations pourront être fournies aux entreprises en question (hôtels, restaurants etc.). Un grand défi consiste surtout en l'investigation multilingue des langues choisies : l'espagnol, l'anglais, le français, l'allemand, le hollandais et l'italien. Le portail de tourisme en ligne Tripadvisor a pu être recruté pour le projet. Ce portail va mettre à disposition la plupart des données pour l'analyse. Toutes les techniques et outils qui seront développés dans ce projet, seront mis à disposition comme outils open source à la fin du projet.

2.5 Conclusion

Dans son département *Human Speech and Language Technologies*, l'entreprise basque Vicomtech est impliquée dans plusieurs projets de recherche différents et ainsi elle couvre beaucoup de domaines de la technologie du langage actuelle. Ce n'est pas seulement la diversité des sujets, mais aussi la collaboration avec beaucoup d'autres entreprises, institutions et projets européens qui font de Vicomtech une entreprise intéressante et convaincante. Un autre trait positif est par ailleurs la forte intégration d'étudiants doctorants et de diplômés universitaires. De plus, Vicomtech joue un rôle important dans le développement de technologies du langage pour la langue basque. Des systèmes tels que la production automatique de sous-titres, ou la synchronisation automatique contribuent à stabiliser le statut de la langue basque. Le plurilinguisme du Pays basque et la promotion de la langue profitent clairement des exploits de Vicomtech dans le domaine de la technologie du langage.

6 http://cordis.europa.eu/newsearch/index.cfm?page=docview&collection=EN_PROJ&reference=103902&position=1&Highlights=OPEN&prevPage=resultList&similarity_id=2633288

3 Mondragon Lingua à Arrasate-Mondragón (04.06.2012)

3.1 Mondragón et la Mondragon Corporation Cooperativa

Mondragón (basque: Arrasate) est une petite ville (à peu près 20.000 habitants) dans la province basque de Gipuzkoa et est située à environ 70 km de San Sebastian à l'intérieur du pays.

En 1943, le prêtre José María Arizmendiarrieta a jeté les bases de la grande réussite de la ville, en fondant une haute école spécialisée, organisée démocratiquement. Après que cette haute école avait produit de nombreuses coopératives prospères, la Mondragón Corporación Cooperativa (MCC) à été créée, dont le principal objectif reste jusqu'aujourd'hui la création systématique d'emplois. La MCC compte actuellement plus de 100 entreprises de trois secteurs très différents: le secteur financier (par ex. banques, compagnies d'assurance), le secteur industriel (comme l'industrie automobile) et le commerce de détail (chaînes de supermarchés). En 1997, la MCC a même construit sa propre université privée à Arrasate-Mondragón, qui porte le nom basque de Mondragon Unibersitatea. Aujourd'hui, elle comprend quatre facultés (sciences humaines, de la nature, d'ingénierie, d'économie et sciences de la restauration) avec environ 4000 étudiants. Un grand avantage de l'université réside dans le fait que les programmes axés sur le marché peuvent grandement bénéficier de la proximité de leur établissement avec le marché du travail.

3.2 Mondragon Lingua

En 1973, le Centre de langues CIM (Centro de Idiomas Mondragón), qui est maintenant appelé Mondragon Lingua, a été fondé dans le secteur des services (qui fait partie du secteur industriel de la MCC). Depuis ce moment-là, l'entreprise a rapidement acquis une réputation comme une école de langues qui a déjà incorporé les technologies de l'informatique pour soutenir les cours de langue au début des années 80. Elle est donc considérée comme une pionnière dans l'utilisation des ordinateurs pour l'enseignement des langues. L'entreprise compte actuellement cinq centres dans le Pays basque : à Bilbo-Bilbao, Donostia-San Sebastián, Arrasate-Mondragón, Gasteiz-Vitoria et à Oñati-Oñate. En plus de l'enseignement des langues, Mondragon Lingua offre également des services de traduction et d'interprétation. Actuellement, plus de 100 traducteurs hautement qualifiés et des interprètes qui traduisent en plus de 40 langues travaillent pour Mondragon Lingua.

3.3 Mondragon Lingua et traduction automatique

Actuellement, Mondragon Lingua poursuit un projet qui va intégrer la traduction automatique (TA) pour la paire de langues anglais-espagnol dans le processus traditionnel de la traduction humaine. Les responsables du projet sont Ane Ruiz de Zarate et Manuel de la Pascua. Les données de formation comprennent, selon de la Pascua, "plusieurs 10 millions de mots" et le système TA (machine translation system) fera usage de Moses. Puisque le projet est encore à ses débuts, il n'existe encore aucune information concrète sur la procédure exacte ou sur la réussite du projet.

Au cours de la présentation du projet par les deux directeurs, il a été clairement démontré que l'usage de la TA est encore marquée d' une certaine réticence (citation de présentation PowerPoint: «MT – because we have no other option»). Toutefois, on arrive à une conclusion positive de la visite chez Mondragon Lingua: On constate que le secteur de la traduction reconnaît de plus en plus la nécessité de la traduction automatique et qu'il saisit l'occasion progressivement.

4 IXA Research Group à Donostia-San Sebastián (05.06.2012)

L'université du pays basque (en espagnol *Universidad del País Vasco*, en basque *Euskal Herriko Unibersitatea*) compte actuellement environ 45 000 étudiants et 5000 collaborateurs académiques[7]. Les 100 départements de l'université sont répartis sur quatre sites : Bilbao, Vitoria-Gasteiz, Eibar et San Sebastián. L'université est bilingue : environ 35 % des étudiants font leurs études en basque.

Le IXA Research Group[8], dont le siège se trouve à San Sebastián, fait partie du département des langues et systèmes informatiques. Le groupe a été fondé en 1988 avec le but de développer des ressources basiques de technologies du langage pour le basque. Ce domaine de recherche porte aujourd'hui le nom plus général de *Natural Language Processing*. A l'heure actuelle, le groupe emploie une équipe interdisciplinaire de 33 informaticiens, 11 linguistes et 3 assistants techniques. De plus, il y a 28 professeurs de l'université qui sont impliqués dans plusieurs projets de IXA. 22 thèses de doctorat sont encadrées par IXA et 8 étudiants vont terminer leurs études d'ici peu. IXA collabore avec des organisations comme Vicomtech ou la fondation Elhuyar, ainsi qu'avec les universités de Barcelone, Amsterdam, Rome et Darmstadt.

4.1 L'activité de recherche

Dans une conférence introductive, Arantza Díaz de Ilarraza nous a donné un aperçu des activités de IXA. Le groupe de recherche s'intéresse à des problèmes généraux du traitement de la langue, en particulier en ce qui concerne le basque. Elle cherche à promouvoir stratégiquement la technologie de la langue pour le basque, en créant des bases et des ressources linguistiques (théoriques) – par exemple des corpus – qui permettent le traitement automatique. Partie du programme font également des outils pour le développement continuant ou même des applications complètes. La liste des produits manufacturés est considérable et couvre des domaines tels que Information Retrieval et l'extraction d'information, la traduction automatique, des logiciels d'apprentissage linguistique, la morphologie, la (morpho-) syntaxe, la lexicographie et la sémantique.

Un produit phare de l'IXA est le vérificateur d'orthographe basque Xuxen (ce qui signifie à peu près 'exact, correct'). Il est disponible gratuitement pour des programmes de traitement de texte (MS Office, OpenOffice) et des produits Mozilla (navigateur, client e-mail) ainsi que pour une application en ligne[9]. Il convient de souligner l'importance d'un vérificateur pour l'orthographe basque: Beaucoup de bascophones manquent d'assurance dans l'orthographe, parce qu'une langue écrite standard n'existe que depuis 1966[10] et parce que la langue s'est déplacée dans le système éducatif seulement vers l'an 1980.

---

7 www.ehu.es
8 ixa (prononciation [iʃa]). Il s'agit du nom basque pour la lettre X.
9 www.xuxen.com (seulement en espagnol e basque).
10 Comparez la documentation de l'IXA Research Group du 05/06/2012, transparent 9.

IXA a été impliqué dans la création de plusieurs vastes ressources digitales. Pour fabriquer le correcteur d'orthographe, on a établi la EDBL (Euskararen Datu-Base Lexikala, «Base de données lexicale de la langue basque»), qui plus tard pourrait être utilisée pour d'autres applications (comme l'analyse morphologique ou la lemmatisation). Elle couvre environ 100'000 lemmes avec des informations morphologiques. Une autre ressource lexicale est le WordNet basque BasWN, une base de connaissances qui met en relation structurée et lexico-sémantique le sens des mots. Elle contient près de 60 000 mots avec un total de plus de 90 000 significations (Synsets).

Deux corpus basques ont également vu le jour: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa, «corpus de référence pour le traitement de la langue basque») est un corpus de 300'000 mots, qui contient l'annotation manuelle sur la morphologie, les phrases, les dépendances syntaxiques, les rôles sémantiques et les Synsets de BasWN. Le corpus plus récent, seulement annoté morpho-syntaxiquement, le ZTC (Zientzia eta Teknologiaren Corpusa, «corpus des sciences et de la technologie»), est beaucoup plus grand avec 10 millions de mots; environ un cinquième des annotations générées automatiquement a été corrigé manuellement. Une interface en ligne gratuitement accessible[11] permet des requêtes complexes au ZTC.

La dimension du corpus a mené à la question collective comment cette quantité de textes scientifiques dans la langue basque pourrait être mise à disposition et s'il existe vraiment des publications en basque. Les publications technologiques ou d'autres publications d'intérêt international sont écrites majoritairement en anglais, celles d'intérêt national souvent en espagnol. Mais des publications spécialisées qui rejoignent un auditoire basque existent aussi. L'université, qui offre une partie considérable des cours parallèlement en espagnol et en basque, a donc fortement besoin de la littérature basque. Cependant, il n'y avait pas assez de matériel digital à disposition pendant longtemps: Pour le corpus, beaucoup de matériel imprimé a été digitalisé.

4.2 KYOTO – Knowledge Yielding Ontologies for Transition-based Organization

Aitor Soroa nous a rapproché du projet KYOTO, un système text-mining polyglotte pour le domaine de l'environnement et protection de la faune. Le projet financé par l'UE a pour objet une plateforme ouverte et librement accessible pour la connaissance liée à l'environnement. Le projet est dirigé par professeur Piek Vossen de l'Université libre d'Amsterdam et comprend des groupes de recherche en Italie, Espagne, République tchèque, la République de Chine, le Japon et l'Allemagne.

L'idée est que le système de text-mining trouve automatiquement les événements (les faits) dans des textes de langues différentes et qu'il les mette à disposition à travers les langues et les cultures.

Il prend en charge les langues anglaise, espagnole, hollandaise, italienne, chinoise et japonaise. A cet égard, WordNet fonctionne comme pont linguistique. Les WordNets non-anglais (voir EuroWordNet[12]) se basent structurellement sur le WordNet anglais et se relient à celui-ci. Ainsi, les

11 www.ztcorpusa.net/cgi-bin/kontsulta.py (seulement en basque).
12 www.illc.uva.nl/EuroWordNet/.

différentes expressions peuvent se référer à des concepts pertinents.
Dans une première étape, le système crée une base de données terminologique. Les soi-disant 'TyBots' (Term Yielding Bots) extraient des termes à l'aide de Wikipedia, ils les projettent sur des concepts et construisent une ontologie.
La propre extraction de connaissances est réalisée avec des centaines de règles écrites à la main. Les soi-disant 'KyBots' (Knowledge Yielding Bots) génèrent un modèle de sortie pour chaque phrase quand une règle est remplie. Les règles contiennent toujours un composant morphosyntaxique et une condition sémantique à la fois, qui tous les deux doivent être remplis. Celle-ci est la tâche la plus difficile, puisque la gamme des expressions de la langue naturelle doit être prise en compte.

La connaissance extraite se présente finalement à trois niveaux: lexical (termes), WordNet (concepts) et dans l'ontologie. Les faits extraits sont stockés dans un format XML (KAF, Kyoto Annotation Framework), qui est approprié pour un le traitement sémantique ainsi que linguistique et qui est flexible et extensible.

4.3 OpeNER – Open Polarity Enhanced Named Entity Recognition

Rodrigo Agerri nous a informé sur le projet OpeNER. Le projet est une collaboration avec Vicomtech et est décrit plus en détail au chapitre 2.5.

4.4 Traduction automatique

Gorka Labaka nous a informé à propos des activités dans la traduction automatique du groupe IXA. Le système Matxin que certains d'entre nous connaissions déjà, a été développé ici. Il est basé sur le corpus et les règles et prend en charge les paires de langues espagnol-basque et anglais-basque.
Le nouveau système TA hybride SMatxinT va maintenant porter les avantages de la TA statistique dans les traductions. TA statistique diffère de la TA basée sur les règles normalement par une meilleure couverture lexicale et la sélection ainsi que par une meilleure manipulation de structures inconnues.
Comparé aux systèmes TA statistiques pour la plupart des langues de l'Europe occidentale, le basque comme langue agglutinante et donc extrêmement riche en formes présente des difficultés supplémentaires. Grace à l'inflexion riche, le basque a un débit très fort de « types» et « tokens », ce qui conduit à une austérité dans les données d'entraînement. Ce qui complique les choses, c'est que la disponibilité de données parallèles pour le basque est limitée. En même temps, le basque contient généralement moins de tokens par phrase que, par exemple, la traduction en espagnol, ce qui conduit à de fréquents alignements 1:n. Ceux-ci tendent à influer sur l'exactitude de la traduction, puisque la généralisation est plus petite.
Ce problème peut être désamorcé par un système de morphologie qui segmente les mots et qui

les divisée en radicaux et affixes ou des groupes d'affixes. Par un choix judicieux de la distribution, c'est possible d'aligner des affixes verbaux du basque avec des verbes auxiliaires de l'espagnol, ce qui augmente la quotepart des alignements 1:1 et réduit le nombre de types. Ainsi, aussi les mot-formes qui n'apparaissaient pas dans le matériel d'entraînement peuvent être traduits correctement.

Le système hybride SMatxinT utilise comme échafaudage pour la traduction un sous-système fondé sur des règles. Avec des règles syntaxiques la structure de la phrase cible est spécifiée. Ensuite, le sous-système statistique calcule pour chaque phrase ou morceaux de phrase plusieurs candidats de traduction alternatives à ceux du système fondé sur des règles, parmi lesquels un est choisi.

4.5. Analyse sémantique

L'après-midi, une brève introduction aux collections de données principales et plus pertinentes pour les projets d'IXA a eu lieu. Ci-dessous se présentent quelques-unes des ressources examinées cet après-midi.

General Inquirer (1966)

Le projet General Inquirer[13] est décrit sur son site web comme une approche informatique pour l'analyse du contenu des textes et comme un outil du mapping. Rodrigo Aguerri a déclaré qu'il est une ressource pour les opinions, dans laquelle les mots inclus sont classifiés selon des valeurs positives ou négatives. Concrètement, cela signifie que tout fichier texte sera mappé selon des catégories. Ce total de 182 catégories se compose d'une liste de mots et de significations. La catégorie «négatif» est, selon le projet, la plus grande avec 2291 entrées. Les utilisateurs ont la possibilité de créer leurs propres grandes catégories arbitraires. L'outil est monolingue, à savoir disponible uniquement en anglais. Il s'agit d'une ressource qui a été utilisé dans la recherche fondamentale de IXA. [14]

EuroWordNet

Le EuroWordNet[15] avait pour but de créer une base de données multilingue avec WordNets pour plusieurs langues européennes. Il a été achevé en 1999. Les langues utilisées sont le néerlandais, italien, espagnol, allemand, français, tchèque et estonien. Comme une multitude de projets WordNet, EuroWordNet s'oriente vers le Princeton WordNet pour l'anglais américain. Tous les WordNets sont conçus de manière qu'ils peuvent être liés avec ceux d'autres langues à travers une catégorie de synonymes (Synsets). Chaque langue a, cependant, son système spécifique pour la lexicalisation. La combinaison de différents dictionnaires de langue se fait par le soi-disant Inter-Lingual-Index (par des synonymes, relations sémantiques). En EuroWordNet, il ya aussi une « top-ontology » composée de 63 distinctions sémantiques que toutes les langues ont en commun. EuroWordNet a été présenté comme un exemple de base d'un WordNet, qui a servi de source

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

d'inspiration et de point de départ pour les propres projets de l'IXA, tels que le WordNet Basque.

SentiWordNet

Des différentes versions de SentiWordNet[16] (1,0 et 3,0) existent. Fondamentalement, il s'agit d'une ressource lexicale pour « opinion mining ». Le SentiWordNet charge chaque Synset d'un WordNet avec trois valeurs de sentiment: positivité, négativité et objectivité. Pour la recherche de l'IXA, le SentiWordNet 1,0 du 2006 est particulièrement pertinent.

WordNet-Affect

WordNet-Affect[17] est une extension d'un autre projet WordNet: WordNet Domains. Comme son nom l'indique, ce WordNet contient une catégorie supplémentaire de Synsets, qui sert à illustrer des concepts et des mots affectifs. Ces Synsets de WordNet obtenaient des étiquettes affectives (a-labels) qui ont été appliquées à des concepts affectifs.

Q-WordNet

Le Q-WordNet[18] est l'un des projets (en cours) de Rodrigo Agerri qui traitent de l'analyse des sentiments. Il peut être téléchargé sur son site personnel. Ce WordNet est une ressource lexique, consistant des significations WordNet qui ont été classées automatiquement avec la polarité positive ou négative. Le classement de polarité est principalement utilisé pour faire des analyses de sondages, tels que le projet précédemment introduit OpeNER. Q-WordNet est inspiré par SentiWordNet. À la différence à SentiWordNet, l'information linguistique n'est pas seulement annotée manuellement mais aussi à la main. Pour Rodrigo Agerri et son équipe, ceci signifiait une dépense supplémentaire, mais aussi une amélioration qualitative importante du matériau.

4.6 Bilan

Le groupe IXA attache beaucoup d'importance au multilinguisme. Comme Université du Pays basque, le basque est une préoccupation naturelle de l'éducation et de la recherche. Le basque est activement promu non seulement aux activités de formation, mais aussi par des projets de recherche tels que le programme de vérification orthographique Xuxen et la création de ressources numériques (EDBL, BasWN, EPEC et ZTC). Le système de traduction automatique SMatxinT apporte également une contribution à combler les barrières linguistiques à travers les deux paires de langues espagnole-basque et anglais-basque. Les projets en cours tels que KYOTO et OpeNER élargissent également l'orientation multilingue de l'IXA au-delà des langues de l'Europe occidentale.

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

5 Le musée San Telmo à Donostia-San Sebastián (06.06.2012)

Le musée San Telmo à San Sebastián, inauguré en 1902, est le plus ancien musée du Pays Basque. Au début, il était consacré aux domaines de l'art, de l'histoire et de l'archéologie; un quatrième domaine, celui de l'ethnographie, s'y est ajouté ensuite pour répondre à une demande croissante.[19] Le but principal du musée est de promouvoir l'héritage historique et socio-culturel du Pays Basque.

Une étape importante dans l'histoire du musée fut sa rénovation et son extension. Depuis 1932, les expositions étaient hébergées dans le monastère dominicain de San Sebastián, un imposant bâtiment qui fut érigé au pied de la forteresse Urgull. L'expansion du musée et l'état précaire du monastère ont mené à la rénovation et à la construction d'un nouveau bâtiment, un projet qui fut dirigé par les architectes espagnols Nieto et Sobejano. Rénové et restructuré, le musée se nomme, depuis sa réouverture en 2011, « Musée de la société basque et de la citoyenneté » (Museo de Sociedad Vasca y Ciudadanía).

L'exposition permanente s'articule en cinq catégories, qui permettent d'entreprendre un voyage à travers l'histoire de la culture et de la société basque : le développement historique du site (historia del sitio), les enjeux de notre société (los desafios de nuestra sociedad), les plus importantes étapes dans l'histoire basque (huellas en la memoria), l'éveil de la modernité (despertar de la modernidad), et la collection historique d'art (collección histórica de arte).

Comme on l'a déjà indiqué, l'ampleur des collections du musée San Telmo n'a fait que croître depuis sa fondation. Il en est de même pour la demande et l'intérêt des visiteurs provenant d'Espagne, d'Europe, et de plus en plus d'autres régions du monde, y compris de Chine et de Russie. Le musée fait donc face, depuis plusieurs années, au défi d'organiser son contenu le mieux possible et de la manière la plus accessible. Pour des raisons d'espace et de coût, la décision a été prise de limiter les langues pour la documentation et les descriptions dans les salles d'exposition à deux seulement: l'espagnol et le basque, les langues officielles du Pays Basque.

Cette initiative a été saluée par la population locale et par les institutions éducatives.

D'une part, la langue basque est considérée comme une composante importante de la culture basque; d'autre part, depuis l'introduction de la forme standard de la langue basque en 1966, l'enseignement en basque est de plus en plus favorisé. Le matériel pour les expositions est souvent composé en collaboration avec d'autres institutions, d'abord en basque et ensuite traduit en interne vers l'espagnol (ou vice versa).

---

19 Certaines informations proviennent du site internet www.santelmomuseo.com et du document « SAN TELMO MUSEUM.doc », mis a disposition par les collaborateurs du musée.

fournit le matériel multimédia, comme les guides audio ou des court-métrages, mais aussi avec une entreprise de traduction, chargée de traduire les textes de l'espagnol ou du basque vers le français et l'anglais. Les traductions sont fournies manuellement par des traducteurs indépendants et sont contrôlées par les collaborateurs du musée avant d'être publiées. Les textes du site internet sont également mis à jour et traduits manuellement. Le musée justifie son choix du français et de l'anglais de la façon suivante: l'anglais est la langue internationale et la plus parlée au monde, le français a été choisi en raison du nombre élevé de visiteurs du pays voisin qu'est la France. Le musée San Telmo est cependant souvent critiqué du fait que les textes dans les salles d'exposition sont présentés uniquement en espagnol et en basque. De plus, la demande de traductions en d'autres langues, comme l'italien, le russe et le chinois, est en hausse. Le département de communication est donc actuellement occupé à produire du matériel supplémentaire (en forme de dépliants et de guide audio) en d'autres langues pour l'été prochain. Le défi principal consiste à trouver et appliquer une forme de production économique et, surtout, rentable. On est par exemple en train de considérer d'ajouter l'allemand ; cependant, des doutes subsistent, étant donné le nombre de visiteurs germanophones, quant à la rentabilité de cet ajout.
En 2016, Donostia-San Sebastián sera la capitale européenne de la culture. Pour cette occasion au plus tard, le musée San Telmo élargira son offre en services multilingues. Actuellement, le musée est en contact avec une entreprise pour développer un logiciel pour smartphones qui pourrait éventuellement remplacer les guides audio et atteindre un plus grand nombre de visiteurs.
Auparavant, un autre projet important va être réalisé à l'automne 2012, l'exposition temporaire sur la formation et le développement de la langue basque.

6 Le groupe EITB à Bilbo-Bilbao (07.06.2012)

Jeudi, nous avons visité la EITB Broadcasting Company à Bilbao. EITB désigne *Euskal Irrati Telebista ;* il s'agit du premier groupe de radio et télévision basque. Avec ses produits différents de l'information et du divertissement, le groupe atteint plus d'un million de citoyens par jour. La fondation d'EITB a été votée en 1982 par le parlement basque. Peu de temps après, la première chaîne radio était déjà en onde et la chaîne de télévision a également bientôt démarré. Le financement du groupe EITB se fait de 80 % par le gouvernement basque et les 20 % restants se font par la publicité.

Le but d'EITB est la promotion et la diffusion de la langue et culture basque. Cela va de pair avec la standardisation du basque et la réforme de l'éducation. L'usage de la langue basque en public était longuement interdit pendant la dictature de Franco. Dans les régions urbaines comme Bilbao, le basque était remplacé presque complètement par l'espagnol, ce qui a fait que la génération de cette époque ne sait plus parler le basque. La compréhension dans le pays basque est rendue encore plus difficile par les différents dialectes basques. Mis à part la standardisation et la réintroduction du basque comme langue officielle, ce sont aussi les médias qui ont aidé à rediffuser la langue basque. Cela concerne aussi le groupe EITB. Un problème de la réintroduction se trouve dans la différence entre le basque standard et le basque régional : les générations plus jeunes ont uniquement appris le basque standard à l'école et ne l'utilisent presque pas au quotidien. Par contre, les habitants des régions rurales, dont le basque est la langue maternelle, considèrent le basque standard comme une langue artificielle. Le groupe EITB espère de pouvoir compenser cette différence en promouvant le basque.

6.1 Les chaînes du groupe EITB

Aujourd'hui, le groupe EITB entretient 5 stations de radio, 5 chaînes de télévision et depuis 2008 il existe aussi un site web multilingue (www.eitb.com), disponible en basque, espagnol, anglais et français. Les chaînes de télévision ETB-1 et ETB-2, diffusées entièrement en basque et en entièrement en espagnol respectivement , sont en onde depuis les années 80. Deux chaînes internationales sont en outre disponibles par satellite : ETB Sat pour l'Europe et Canal Vasco pour l'Amérique. Toutes les deux transmettent en espagnol et servent à la diffusion de la culture basque, mais elles transmettent aussi des événements sportifs. La chaîne plus récente ETB-3 veut atteindre une audience plus jeune et transmet des contenus en basque et en espagnol. Mais la chaîne prévoit aussi de diffuser des contenus en anglais. En Espagne, traditionnellement, tous les programmes importés sont synchronisés, mais sous l'effet de la globalisation, le groupe EITB brigue de plus en plus aussi la promotion de la langue anglaise. Les chaînes de télévision diffusent sans arrêt ; il existe en effet un accord avec le gouvernement de ne jamais émettre d'images noires.

Le groupe EITB entretient en tout 5 stations de radio : deux chaînes générales, *Euskadi Irratia* en

basque et la radio *Euskadi* en espagnol. *Gaztea* transmet en basque pour l'audience jeune, alors que la radio *Vitoria* est chargée de la région d'Alava. Finalement, il y a encore *EITB Musika*, une station de radio qui transmet seulement de la musique.

6.2 Le travail au siège principal

Le groupe EITB possède des sites dans trois villes : A San Sebastián et à Vitoria se trouve la production des émissions de télévision et des séries télévisées. A Bilbao se situe le siège principal avec les rédactions de radio, de télévision et d'Internet, ainsi que des studios et d'autres endroits de production. L'architecture du siège principal, inauguré en 2007, veut refléter la philosophie du groupe avec un design ouvert et aéré. En même temps, le design symbolique vise aussi à centraliser le processus de travail. D'abord on arrive au secteur de contrôle, qui est le cœur des chaînes de télévision. Dans cette passerelle, tous les contenus sortants et entrants sont bottelés, et on contrôle qu'il n'y ait pas des interruptions dans le programme. Derrière se trouve le 'cerveau' du groupe, les ordinateurs, qui s'occupent d'éditer, vérifier et enregistrer les contenus. La vérification et l'intégration des donnés méta sont très importants pour que les journalistes puissent retrouver les contenus. De plus, tous les contenus sont archivés sous forme digitalisée. Si un contenu n'est pas diffusé sous quatre jours, il est supprimé définitivement. Cela sert à économiser de l'espace mémoire. Si le contenu est utilisé, il est enregistré dans un autre endroit. Le groupe EITB s'engage à enregistrer tout ce qui était transmis pendant quatre ans pour des raisons juridiques. En outre, il entretient un département pour la digitalisation des contenus plus anciens, qui sont aussi stockés dans les archives digitaux.

En-dessous et derrière le secteur de contrôle se trouve un grand espace ouvert avec des postes de travail pour les journalistes. Les postes permettent d'effectuer des recherches, mais aussi le montage et l'édition du son. L'intention initiale était de créer un newsroom combinant les rédactions de radio, télévision et Internet, mais on a bientôt constaté que les processus de travail des trois sections sont trop différents. Ainsi, les rédactions aujourd'hui sont séparées spatialement. Malgré cela, le design ouvert vise justement à diminuer les barrières entre les employés et les chefs.

En outre, il y a aussi plusieurs studios de radio et de télévision, ou se déroule la production des émissions en direct. Celles-ci sont montées en direct et aussi sous-titrés en simultané, à cause d'une nouvelle législation, qui édicte des sous-titres pour les personnes malentendantes. Cela mène bien sûr à la génération de nouvelles données, qui sont très intéressantes pour l'analyse multilingue de textes. C'est pourquoi la production de sous-titres et la digitalisation des contenus plus anciens représentent une passerelle vers la technologie du langage et offrent un point de départ pour l'amélioration les processus de production.

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

Mixer Servicios Audiovisuales S.L. a été créée en 2004 par la fusion des entreprises *Edertrack* et *Irusoin*[20] . Aujourd'hui elle compte parmi les entreprises les plus importantes pour la synchronisation et la production de sous-titres dans la région basque espagnole. Cette entreprise possède deux sièges sociaux : l'un à Bilbao et l'autre à San Sebastián. 32 techniciens hautement qualifiés travaillent chez Mixer, et plus de 90 traducteurs et acteurs indépendants sont employés pour la synchronisation. De plus, Mixer est un membre co-fondateur de l'établissement *EIKEN*, le groupe audiovisuel basque qui entretient une coopération avec les entreprises basques les plus importantes dans le domaine audiovisuel.

7.1 Services

Mixer a beaucoup d'activités différentes, comme la production de sous-titres pour des films et des programmes de télévision, ainsi que la synchronisation. Un autre domaine d'activité est celui de la publicité : Mixer fournit des traductions, des enregistrements et de la musique pour la publicité, pour le E-Learning et pour l'enseignement. Mixer produit aussi des guides audio basques et quelquefois espagnols pour des musées, par exemple pour le musée Guggenheim à Bilbao. Elle s'occupe aussi de l'adaptation du ton et de la musique de films et de différents programmes de télévision. Comme c'est une entreprise basque, la majorité de ses activités se concentre sur la langue basque.

7.1.1 Synchronisation

Une grande partie des productions de l'entreprise est la synchronisation de films, de programmes de télévision et de dessins animés, surtout pour le service de radiodiffusion basque *Euskal Irrati Telebista* (EITB). D'habitude la synchronisation est faite de l'anglais vers le basque ou de l'espagnol vers le basque ; quelquefois aussi de l'anglais vers l'espagnol. Mixer dispose d' archives avec des enregistrements de différents speakers, que l'on peut écouter sur le site web. En coopération avec Vicomtech on a essayé d'adapter et de manipuler la voix des speaker en la transformant, pour que l'on puisse ainsi utiliser le même speaker pour différents rôles. Cependant, on n'a pas réussi à développer un outil ou une solution satisfaisante jusqu'à présent.

La synchronisation de films, DVDs et de programmes de télévision diffère fondamentalement de la production de sous-titres. Ni la traduction automatique, ni des systèmes de mémorisation des traductions peuvent être utilisés pour la production des notes de transcription. Pour la synchronisation, les notes et les mouvements de la bouche doivent être coordonnés. De plus, il est souvent difficile de transférer des jeux de mots, des plaisanteries et des métaphores d'une languedà l'autre sans changer le sens. Un traducteur spécialisé en synchronisation travaille environ 1 à 2 semaines pour synchroniser un film. Le traducteur reçoit la transcription du film dans

20 www.irusoin.com

la langue originale ainsi que le film, afin de pouvoir coordonner le texte de la synchronisation et les mouvements de la bouche. Pour faire cela, il détermine la codification du temps des répliques et les intègre dans les notes. A partir de cela, le film est synchronisé au studio avec les acteurs ; finalement l'éditeur révise le résultat avant la mise en onde.

7.1.2 La production des sous-titres

La production de sous-titres est aussi une des tâches principales de l'entreprise Mixer. La majorité des sous-titres produits sont la transcription de la langue originale du film pour les malentendants. Seulement la minorité des sous-titres sont des traductions de la langue d'orignine. Comme dans le cas de la synchronisation, *EITB* est le client plus important de *Mixer.* Cette année *EITB* a diffusé 75% de ses programmes avec des sous-titres dans la même langue. L'année prochaine devront être produits des sous-titres pour environ 90% des émissions. Contrairement à la synchronisation, beaucoup des sous-titres sont produits semiautomatiquement avec des programmes comme *SuSa* (*semi-automatic subtitling*). *SuSa* a été développée en coopération avec Vicomtech et est utilisé pour la conversion de la transcription de synchronisation en sous-titres. Avec la codification du temps, *SuSa* peut faire une synchronisation des notes et des fichiers audio. La personne qui prépare les sous-titres fait le transfert des notes en sous-titres et en même temps il voit le film. Les sous-titres qu'il ajoute reçoivent la codification du temps et le formatage correspondant. A partir de cela on peut ajouter des informations supplémentaires, par exemple des sons ou des bruits, pour les malentendants.

7.2 Bilan

La visite chez l'entreprise Mixer Servicios Audiovisuales S.L. à San Sebastián a été très variée et orientée vers la pratique. Ils ont non seulement présenté leurs produits, mais aussi montré les différents outils et programmes qu'ils utilisent pour produire des sous-titres. De plus, nous avons eu la possibilité de voir les studios pour avoir une impression de comment se déroule la production d'un programme de télévision avec sous-titres.

En général nous avons constaté que dans ce domaine il y a encore un grand potentiel pour la traduction automatique. Etant donné que la syntaxe du basque diffère forement de celle espagnol et anglaise, et parce que seulement peu de textes parallèles sont à disposition pour ces langues, il est pour l'heure difficile de traduire les sous-titres automatiquement, comme on l'a fait par exemple à l'Université de Zurich, en coopération avec une entreprise suédoise qui produit des sous-titres. Comme l'objectif principal de Mixer est de produire des sous-titres pour les malentendants, cette approche devrait être adaptée différemment. Pour l'avenir ce serait désirable pour l'industrie de pouvoir produire les sous-titres automatiquement ou d'avoir la possibilité d'adapter les transcriptions de la synchronisation automatiquement aux mouvements de la bouche ; cependant, à l'heure actuelle cela reste impossible, et Mixer devra encore produire et traduire beaucoup des sous-titres manuellement.

8 Elhuyar à Donostia–San Sebastián (08.06.2012)

Elhuyar[21] a été créée en 1972 avec le but d'encourager l'emploi de la langue basque dans la science et dans technologie. En 2002 Elhuyar est devenue une fondation à partir de laquelle sont surgi deux entreprises: L'entreprise *Elhuyar aholkuaritza* et *Eleka Ingeniaritza Linguistikoa*. L'entreprise *Elhuyar aholkuaritza* veut promouvoir l'usage de la langue basque. *Eleka Ingeniaritza Linguistikoa* est une entreprise qui offre des produits et services de la technologie linguistique. Le nom *Elhuyar* provient de deux frères basques, José et Fausto Elhuyar, qui ont découvert un élément chimique qui s'appelle Wolfram. Il s'agit de la découverte scientifique la plus importante du Pays basque.

8.1 Le département pour les services linguistiques

La fondation Elhuyar est subdivisée en plusieurs départements, dont un pour les services linguistiques. L'objectif principal du département pour services linguistique est mettre à disposition des ressources linguistiques, des outils et des services pour la langue basque ainsi que pour d'autres langues (par exemple l'espagnol). Le département pour services linguistiques est à son tour subdivisé en différents sous-départements: un sous-département pour les traductions et la correction, un sous-département pour la création de dictionnaires et un sous-département pour l'investigation et le développement de nouvelles technologies et ressources linguistiques. Les domaines de recherche principaux sont l'extraction terminologique, l'extraction d'informations, la création de corpora, la sémantique, les ontologies et la traduction automatique. Dans ce qui suit quelques domaines de recherche principale son décrits.

8.1.1 La création de corpora

Les corpora de Elhuyar sont créés en partie automatiquement, à partir de l'extraction de textes d''internet, et en partie manuellement. Un corpus très important d'Elhuyar est *ZT Corpusa,* un corpus scientifique et technologique qui a été élaboré en coopération avec IXA. Ce corpus est composé de textes basques scientifiques et techniques, publiés entre les années 1990 et 2002. Ce corpus contient une partie balancée (env. 2 millions de mots) et une partie non-balancée (env. 7 millions de mots). C'est possible de rechercher seulement dans la partie balancée ou dans tout le corpus. Le site Web[22], où l'on peut effectuer des recherches dans le corpus est visité environ 2000 fois par mois, surtouts par des professeurs d'université et par des auteurs de lexiques.

Un autre corpus basque d'Elhuyar est *Lexikoaren Behatokia*[23]. C'est un corpus de monitorage, c'est-à-dire un corpus qui est complété régulièrement avec des textes actuels. Les textes proviennent de différents médias et sont mis à disposition par l'académie royale pour la langue

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

basque. Environ 10 millions de mots sont ajoutés par an. Il existe en outre le corpus *CorpEus*[24]. Dans ce cas, l'internet forme le corpus. Cet outil consiste en un filtre qui détecte des mots basque. Avec ce filtre il est possible de chercher des pages Internet basques qui correspondent à la demande donnée.

Elhuyar a aussi créé un corpus multilingue : le corpus des consommateurs. Il se compose de textes des magazines de supermarché (par exemple des comparaisons de produits) en basque, catalan, espagnol et galicien. Ce corpus a été créé manuellement.

8.1.2 L'extraction terminologique

Elhuyar essaye d'extraire automatiquement des terminologies de corpora et d'ainsi automatiser une partie de la création de lexiques. Pour extraire des terminologies, des méthodes linguistiques et statistiques sont utilisées : il s'agit donc d'un système hybride. Une méthode exclusivement statistique serait inconvenante, parce que le basque est une langue fortement agglutinante, d'où beaucoup de formes de mot différentes de chaque lemme.

Dans le contexte de l'extraction de terminologies, l'outil *Itzul Term*[25] a été développé. Celui-ci peut extraire automatiquement des mots basques et espagnols qui se correspondent à partir des systèmes de mémorisation de traductions. Il se peut que cet outil sera bientôt aussi adapté pour l'anglais. Etant donné que seulement des corpora parallèles pour les langues basque-espagnol et basque-anglais sont à disposition, des expériments avec des corpora comparables sont également effectués. Dans le contexte d'une thèse de doctorat, on essaye en plus d'extraire des phrases idiomatiques de corpora basques.

Une autre possibilité d'automatiser la création de lexiques est la technique Pivot. Si on veut par exemple créer un lexique pour les langues basque-chinois et si on a à disposition des lexiques pour les langues basque-anglais et anglais-chinois, on peut utiliser l'anglais comme langue intermédiaire (c'est-à-dire comme langue pivot). On cherche pour un certain mot basque tous les mots correspondants en anglais et pour tous ces mots correspondants anglais on cherche tous les mots correspondants chinois. Tous ces mots chinois trouvés sont les traductions possibles. A l'aide du contexte, on peut vérifier quels mots trouvés sont vraiment des traductions correctes.

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

9 Une semaine intensive dans le Pays basque - Conclusion

9.1 Le basque - possibilités et limites de la technologie linguistique

Ils existent plusieurs raisons pour lesquelles le basque a une position difficile dans la technologie de la langue.

À cause du caractère isolé de la langue, la linguistique computationnelle, par exemple, ne peut pas bénéficier de la proximité à d'autres langues et ainsi obtenir aucun avantage des paires de langues étroitement apparentées, comme c'est le cas avec des traductions d'une langue source à une langue cible de la même famille de langues.

La morphologie fortement agglutinante de la langue basque apporte également des problèmes fondamentaux, tels que la question de l'unité du mot: Qu'est-ce qu'un mot? Dans le basque, une chaîne (angl. « string ») est normalement constituée d'une variété de morphèmes et doit donc souvent être alignée avec plusieurs mots d'une autre langue moins fortement agglutinante, créant ainsi une correspondance un-à-plusieurs.

Pour aggraver les choses, le basque, qui est composé de sept dialectes principaux parfois très différents, est seulement normalisé depuis 1966, et ce processus n'est pas encore complet. Pour cette raison, la technologie de la langue ne peut pas recourir à un patrimoine exhaustif de la littérature avec un langage homogène, ce qui explique pourquoi les données d'entraînement pour des projets tels que la traduction automatique ne sont pas très vastes. Pour ces fins, par exemple, le groupe IXA s'engage principalement en raison de textes plus récents, tels que livres, journaux de langue ou de la littérature technique. La quantité de données et la sélection des données, cependant, sont sévèrement limitées.

À ces fins, par exemple, le groupe IXA recourt principalement à des textes ultérieurs comme des livres de langue, des journaux ou de la littérature technologique. La quantité et la sélection des données, cependant, sont très limitées.

Le problème des données d'entraînement se présente également dans le projet SUMAT: Bien que Vicomtech est impliquée dans le projet comme entreprise basque, il n'y a pas de traduction des sous-titres de film au basque - sans parler du basque comme langue source. La raison est d'une part le manque de disponibilité des sous-titres existants en basques comme matériel d'entraînement pour le système de traduction, d'autre part, il n'existe même pas la nécessité d'une paire de langues de traduction avec le basque comme langue cible. Comme les locuteurs du basque bénificient presque tous d'une éducation multilingue, ils peuvent donc se servir des sous-titres espagnols.

Bien que la technologie de la langue dans le Pays basque fait un grand effort pour promouvoir la langue basque, les ressources en technologies linguistiques sont encore fortement limitées. Donc, jusqu'à aujourd'hui, ils n'existent que peu de dictionnaires basques et un seul tagger (open-source).

Les visites aux diverses entreprises basques dans le domaine de la technologie de la langue ont montré que les efforts des linguistes computationnels basques ont produit de nombreux projets importants. Ainsi, la Fondation Elhuyar a développé (à côté de beaucoup d'autres choses) un chercheur web qui – supplémentaire à Google - cherche exclusivement les sites et des textes basques sur Internet. Avec BerbaTek, la Vicomtech a notamment construit un système audiovisuel question-réponse pour le basque et le groupe IXA a construit – avec une équipe autour de Gorka Labaka – un propre système de traduction automatique, basé sur des règles (Matxin) pour la paire de langues espagnol-basque. Ce ne sont que trois exemples de nombreux projets prometteurs qui nous ont été présentés cette semaine.

En résumé, on peut dire que la normalisation de la langue basque depuis 1966 a été une chance pour le traitement technologique de la langue. Elle permet que des données d'entraînement normalisées sont créées peu à peu, qui sont capitales pour de nombreux projets de la linguistique computationnelle. Ainsi, les entreprises s'occupent de plus en plus du défi dans le traitement linguistique computationnel de la langue basque et la promeuvent en même temps, en permettant l'accès à la langue, qui s'est adapté aux moyens de communication modernes.

9.2 Comment est-ce-que la technologie linguistique profite de la multilingualité?

Et comment est-ce-que la multilingualité profite de la technologie linguistique?

Dans les années passées le basque est devenu plus important, parce qu'il est devenu une matière obligatoire dans les écoles et parce qu'il est souhaité ou pourvue que les employés sachent le basque. La standardisation du basque (Euskara Batua) a aussi promu l'usage du basque. Dans le pays basque le basque est aujourd'hui promu est propagé, mais d'un point de vue international le basque, au contraire du française, ne joue pas un rôle important. À cause de cela on ne produit pas ou seulement peu des produits et outils pour le basque hors du pays basque. Mais à cause de la reprise du basque l'industrie dans la région basque a besoin des outils techniques et linguistiques pour le basque. À cause de cela et à cause de l'identification forte des basques avec sa culture beaucoup des entreprises focalisent la production des outils pour le basque. En plus la majorité des institutions produit aussi des outils pour l'espagnol et l'anglais.

La technologie linguistique profite de la multilingualité parce qu'on développé des nouveaux outils pour le basque et parce qu'on adapte des outils existantes à la langue basque. L'isolation du basque (chapitre 2.2) et sa complexité linguistique provoquent des nouveaux défis pour les développeurs. Les solutions trouvées pour le basque on peut ensuite utiliser aussi pour autres langues ou pour soluer autres problèmes de la technologie linguistique.

La multilingualité profite aussi de la technologie linguistique. Les outils de la technologie linguistique simplifient le transfert entre l'espagnol et le basque (par exemple par la traduction automatique) et aident chez l'acquisition du basque comme langue secondaire (par exemple si on a des système de correction automatique ou des systèmes pour apprendre une langue). En plus

les outils de la technologie linguistique, comme par exemple les outils de recherche, motivent les gens d'utiliser le basque.

Cette semaine dans le pays basque nous a montré que la technologie linguistique et la multilingualité s'influencent positivement. Nous avons vu que la technologie linguistique s'occupent aussi avec des langues minoritaires. Comme ça d'une parte la technologie linguistique peut profiter et d'autre part on aide de conserver et promouvoir les langues minoritaires.