Valuable Language Resources and Applications Supporting the Use of Basque

Iñaki Alegria, Maxux Aranzabe, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Aingeru Mayor, and Kepa Sarasola

Ixa Group. University of the Basque Country
i.alegria@ehu.es
http://ixa.si.ehu.es

Abstract. We present some Language Technology applications and resources that have proven to be valuable tools to promote the use of Basque, a low density language. We also present the strategy we have followed for almost twenty years to develop those tools and derived applications as the top of an integrated environment of language resources, language tools and other applications. In our opinion, if Basque is now in a quite good position in Language Technology is because those guidelines have been followed.

Keywords: Language resources, Language Technology applications, Strategy for Language Technology development.

1 Introduction

Basque is both a minority and a highly inflected language with free order of sentence constituents. Language Technology for Basque is thus both, a real need and a test bed for our strategy for developing language tools for Basque.

Basque is an isolate language, and little is known about its origins. It is likely that an early form of the Basque language was already present in Western Europe before the arrival of the Indo-European languages.

Basque is an agglutinative language, with a rich flexional morphology. In fact for nouns, for example, at least 360 word forms are possible for each lemma. Each one of the grammar cases as *absolutive*, *dative*, *associative* has four different suffixes to be added to the last word of the noun phrase. These four suffix variants correspond to *undetermined*, *determined singular*, *determined plural* and *close determined plural*.

Basque is also an ergative-absolutive language. The subject of an intransitive verb is in the absolutive case (which is unmarked), and the same case is used for the direct object of a transitive verb. The subject of the transitive verb (that is, the agent) is marked differently, with the ergative case (shown by the suffix *-k*). This also triggers main and auxiliary verbal agreement.

The auxiliary verb, or periphrastic, which accompanies most main verbs, agrees not only with the subject, but with the direct object and the indirect object, if present. Among European languages, this polypersonal system (multiple

Z. Vetulani (Ed.): LTC 2009, LNAI 6562, pp. 327–338, 2011.

verb agreement) is only found in Basque, some Caucasian languages, and Hungarian. The ergative-absolutive alignment is rare among European languages, but not worldwide.

It remains alive but in last centuries Basque suffered continuous regression. The region in which Basque is spoken is smaller than what is known as the Basque Country, and the distribution of Basque speakers is not homogeneous there. The main reasons of this regression during centuries [5] were that Basque was not an official language, that it was out of educational systems, out of media and out of industrial environments. Besides, the fact of being six different dialects made difficult the wide development of written Basque.

However, after 1980, some of those features changed and many citizens and some local governments promote recovering of Basque Language.

Today Basque holds co-official language status in the Basque regions of Spain: the full autonomous community of the Basque Country and some parts of Navarre. However, Basque has no official standing in the Northern Basque Country. In the past Basque was associated with lack of education, stigmatized as uneducated, rural, or holding low economic and power resources. There is not such an association today, Basque speakers do not differ from Spanish or French monolinguals in any of these characteristics.

Standard Basque, called *Batua* (unified) in Basque, was defined by the Academy of Basque Language[1] (Euskaltzaindia) in 1966. At present, the morphology is completely standardized, but the lexical standardization process is underway. Now *Batua* is the language model taught in most schools and used on the few media and official papers published in Basque.

We are around 700,000 Basque speakers, around 25% of the total population of the Basque Country, and we are not evenly distributed. But still the use of Basque in industry and especially in Information and Communication Technology is not widespread. A language that seeks to survive in the modern information society has to be present also in such field and this requires language technology products. Basque as other minority languages has to make a great effort to face this challenge [13,16].

2 Strategy to Develop HLT in Basque

IXA is a research group created in 1986 by 5 university lecturers in the Computer Science Faculty of the University of the Basque Country with the aim of laying foundations for research and development of NLP software mainly for Basque. We wanted to face the challenge of adapting Basque to language technology.

Twenty three years later on, now IXA[2] is a group composed by 28 computer scientists, 13 linguists and 2 research assistants It works in cooperation with more than 7 companies from Basque Country and 5 from abroad; it has been involved in the birth of two new spin-off companies; and there are several products of language technology we have built.

[1] `http://www.euskaltzaindia.net`
[2] `http://ixa.si.ehu.es`

In recent years, several private companies and technology centers of the Basque Country have begun to get interested and to invest in this area. At the same time, more agents have come to be aware of the fact that collaboration is essential to the development of language technologies for minority languages. Fruits of this collaboration were the HIZKING21 project (2002-2005) and ANHITZ project (2006-2008). Both projects were accepted by the Government of the Basque Country as a new strategic research line called *Language Info-Engineering*.

At the very beginning, twenty three years ago, our first goal was to create just a translation system for Spanish-Basque, but after some preliminary works we realized that, being Basque so different from their neighboring languages, instead of wasting our time in creating an ad hoc MT system with small accuracy, we had to invest our efforts in creating basic tools and resources for Basque (morphological analyzer/generator, syntactic analyzers...) that could be used later on to build not just a more robust MT system but also any other language application.

This thought was the seed to design our strategy to make progress in the adaptation of Basque to Language Technology. This way we could face up to the scarcity of the resources and tools, and could make possible the development in Language Technology for Basque at a reasonable and competitive rate.

We presented an open proposal for making progress in Human Language Technology [1]. Anyway, the steps proposed did not correspond exactly with those observed in the history of the processing of English, because the high capacity and computational power of new computers allowed facing problems in a different way.

Our strategy may be described in two points:

1. Need of *standardization* of resources to be useful in different researches, tools and applications
2. Need of *incremental design and development* of language foundations, tools, and applications in a parallel and coordinated way in order to get the best benefit from them. Language foundations and research are essential to create any tool or application; but in the same way tools and applications will be very helpful in the research and improvement of language foundations.

Following this, our steps on standardization of resources brought us to adopt TEI and XML standards and also to the definition of a methodology for corpus annotation [7].

In the same way, taking as reference our experience in incremental design and development of resources/tools, we propose four phases as a general strategy for language processing:

1. Foundations.
   Corpus I (collection of raw text without any tagging mark). Lexical database I (the first version could be just a list of lemmas and affixes). Machine-readable dictionaries. Morphological description.

2. Basic tools and applications.
   Morphological analyzer. Lemmatizer/tagger. Spelling checker and corrector (although in morphologically simple languages a word list could be enough, in Basque we can not take this approach). Speech processing at word level. Corpus II (word-forms are tagged with their part of speech and lemma). Lexical database II (lexical support for the construction of general applications, including part of speech and morphological information). Statistical tools for the treatment of corpus.
3. Advanced tools and applications.
   An environment for tool integration. Web crawler. A traditional search machine that integrates lemmatization and language identification. Surface syntax. Corpus III (syntactically tagged text). Grammar and style checkers. Structured versions of dictionaries (they allow enhanced functionality not available for printed or raw electronic versions). Lexical database III (the previous version is enriched with multiword lexical units, semantic information). Integration of dictionaries in text editors. Lexical-semantic knowledge base. Creation of a concept taxonomy (e.g.: Wordnet). Word-sense disambiguation. Speech processing at sentence level. Computer Aided Language Learning (CALL) systems.
4. Multilingualism and general applications.
   Information retrieval and extraction. Question/Answering. RBMT and SMT Machine Translation System development and Translation aids (integrated use of multiple online dictionaries, translation of noun phrases and simple sentences). Corpus IV (semantically tagged, annotation of senses, argument-structure of sentences). Extraction of information based on semantics. Anaphora resolution and study of discourse markers.

We complete this strategy with some suggestions about what shouldn't be done when working on the treatment of minority languages. a) Do not start developing applications if linguistic foundations are not defined previously; we recommend following the above given order: foundations, tools and applications. b) When a new system has to be planned, do not create *ad hoc* lexical or syntactic resources; you should design those resources in a way that they could be easily extended to full coverage and reusable by any other tool or application. c) If you complete a new resource or tool, do not keep it to yourself; there are many researchers working on English, but only a few on each minority language; thus, the few results should be public and shared for research purposes, for it is desirable to avoid needless and costly repetition of work.

There are other interesting works related to general policies to develop resources and applications for low-density languages [14,9].

3 Useful Applications and Resources

In this section we describe four effective applications and four language resources already created by our group.

3.1 Spelling Checker/Corrector

Because the use of Basque was forbidden during many years in schools and also because of its late standardization[3], adult speakers nowadays did not learn it at school, and so they write it imperfectly. For example, when someone goes to write the word *zuhaitza* (tree), the many possible spellings (*zuhaitz*? *zugaitz*? *zuhaitx*? *zuhaitsa*? *sugatza*?) may cause the writer to hesitate, often leading to an easy solution: *Give up, and write the whole text in Spanish or French!*

The spelling checker Xuxen [2] is a very effective tool in this kind of situation, giving people more confidence in the text they are writing. In fact, this program is one of the most powerful tools in the ongoing standardization of Basque.

The spelling checker is more complex than equivalent software for other languages, because most of those are based on recognizing each word in a list of possible words in the language. However, because of the rich morphology of Basque, it is difficult to define such a list, and consequently, possible morphological analysis must be included. Xuxen is publicly available from `http://www.euskara.euskadi.net`, where there have been more than 20,000 downloads. There are versions for Office, OpenOffice, Mozilla, PC, Mac, and also an online web service (`http://www.xuxen.com`).

The version for Office includes morphological analysis, but, what happens if we want to use the speller in the "free world" (OpenOffice, Mozilla, emacs, LaTeX, ...)? ispell and similar tools (aspell, hunspell, myspell) are the usual mechanisms for these purposes, but they do not fit with the two-level model [11] we have to use to be able to describe Basque morphology. In the absence of two-level morphology, our solution was to adapt the two-level description to *hunspell* in a (semi)automatic way. With the stems and two sets of suffixes, corresponding to the paradigms at first and second level, which have been obtained all the information we needed for the hunspell description was ready. Only a format conversion was necessary for delivery the spelling checker/corrector for OpenOffice, and other tools integrating hunspell (`http://www.euskara.euskadi.net`) In addition, we also adapted the description to *myspell* that is useful for open source programs like Firefox that have not yet integrated *hunspell* (`http://www.librezale.org/mozilla/firefox`); to do this we combine the main paradigms (here with restricted generation power for each one) and the inclusion of the word forms appearing in a big corpus, after eliminating forms rejected by the original complete spelling checker. Although those approaches for the *free world* have lesser coverage for Basque morphology, they are very useful spelling checkers. As a reference of its use we can mention that more than 115.000 downloads[4] have been done since 2007 for this Firefox add-on.

[3] The academy of Basque defined the morphology and verbs of Unified Basque in 1966, but the lexical standardization process is still going on.

[4] `http://addons.mozilla.org/en-US/firefox/addon/4020`

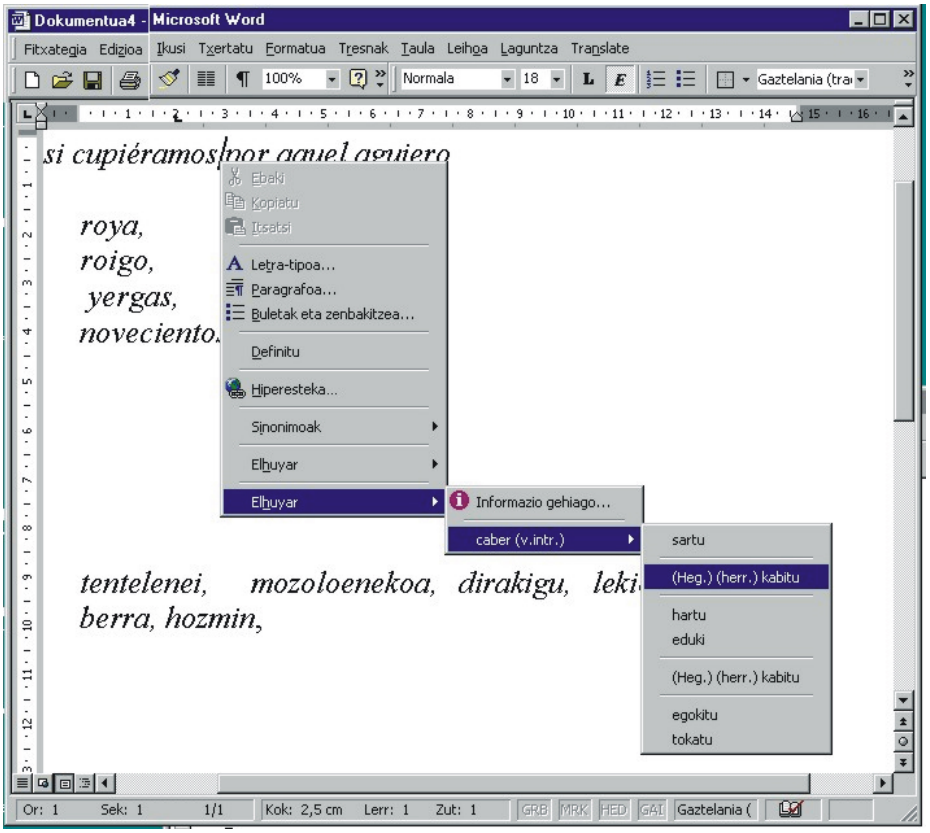

Fig. 1. Lemmatization-based on-line dictionary consulting

3.2 Lemmatization-Based On-Line Dictionaries

The main product created for this kind of application is a plug-in for MS Word that enables looking up a word in several dictionaries; but, in order to make it more useful for a language like Basque with its rich morphology, the dictionary is enhanced with lemmatization. This means that morphological analysis is first performed, and then possible lemmas of the word are matched with the dictionary.

In the example shown in Fig. 1, the user asks for the meaning in Basque of the Spanish word *cupiéramos*. That word-form cant be found in paper dictionaries because it is a finite verb form, but the application recognizes that it corresponds to the verb *caber* (Basque for to fit), and shows five different equivalents in Basque for that verb.

At the moment this plug-in works with three dictionaries: Spanish-Basque, French-Spanish and a dictionary of synonyms. The Spanish-Basque version is publicly available in `http://www.euskara.euskadi.net`.

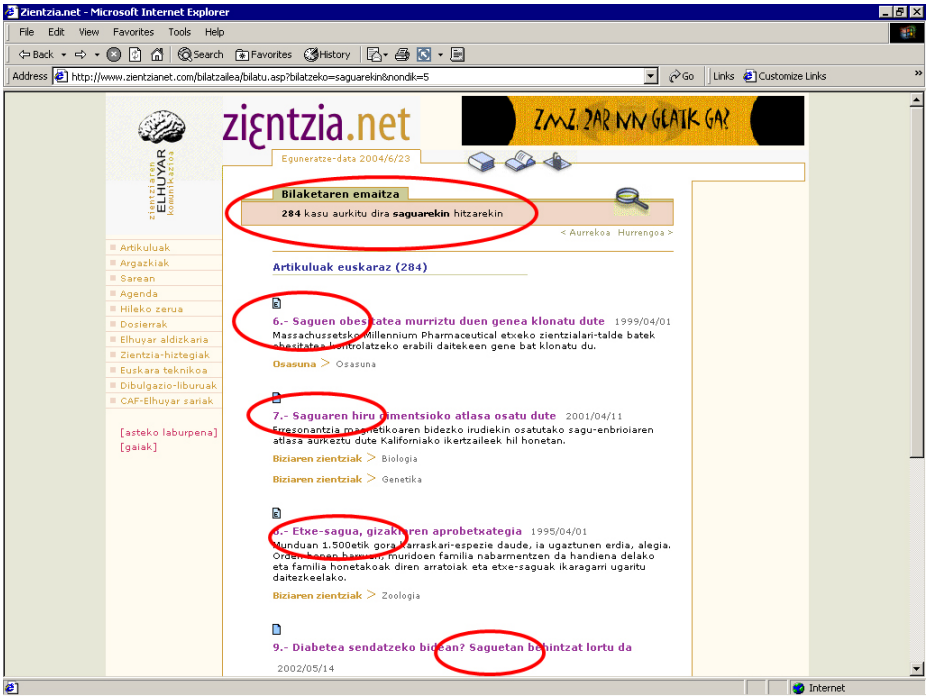

Fig. 2. Lemmatization based document search

3.3 Lemmatization-Based Search Machine

We have developed a search machine to be used with text documents.. This program first performs morphological analysis of the word, and then searches relevant documents containing the lemmas corresponding to these possible morphological decompositions. In the example shown in Fig. 2 the user is searching in the Elhuyar[5] science divulgation journal for documents related to the Basque word form *saguarekin* (with the mouse). The search machine looks for documents containing words whose lemma is just *sagu* (mouse): *saguen*, *saguaren*, *sagua*, *saguetan*...

The principal search machines available nowadays do not have this ability; therefore, if you want to find *sagu* (mouse), you will only find occurrences of that exact word, or alternatively, when searching for any word beginning with that word (*sagu**), many irrelevant documents will be found that contain words such as *saguzar* (bat) which do not correspond to the desired lemma. Consequently, lemmatization-based search machines give users better results.

3.4 Transfer-Based Machine Translation System

When we have faced a difficult task such as Machine Translation into Basque, our strategy has worked well. In 2000, after years working on basic resources

[5] `http://www.zientzia.net`

Fig. 3. Matxin MT system

and tools, we decided it was time to face the MT task. Our general strategy was more specifically defined for Machine Translation, and we had in mind the following concepts:

1. *Reusability* of previous resources, especially lexical resources and morphology description.
2. *Standardization* and *collaboration*: at least, using a more general framework in collaboration with other groups working in NLP.
3. *Open-source*: this means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance it to produce a new MT system, even for other pairs of related languages or other NLP applications.

We have gotten good results in a short time by just reusing previous work, reusing other open-source tools, and developing only a few new modules in collaboration with other groups.[6]

In addition, we have produced new reusable tools and defined suitable formats. We created Matxin using a transfer rule-based MT approach. It translates text from Spanish into Basque, and two results produced in the machine translation track are publicly available:

[6] Opentrad project: `http://www.opentrad.org`

– http://matxin.sourceforge.net for the free code of the Spanish-Basque system and
– http://www.opentrad.org for the online version.

Now we are working in the construction of EBMT and SMT systems and a multi-engine system including three subsystems based on different approaches to MT: rule-based machine translation, statistical machine translation and example-based machine translation [4].

3.5 EDBL: Lexical Database for Basque

EDBL is a general-purpose lexical database of Basque, and so it constitutes an essential foundation for any task in the field of automatic processing of the language. It was first developed as lexical support for the spelling checker, but nowadays it constitutes also the basis for the lexical component of different tools such as a morphological analyzer, a lemmatizer, a multiword lexical units' recognizer, a named entities' recognizer, and so on. It has proved to be a multipurpose resource, from which tailored lexicons can be exported.

Following a mass enrichment process carried out two years ago, its content increased a 25%, thanks to the collaboration with UZEI[7] and Elhuyar[8], in the frame of a corpus project named The Observatory of the Lexicon[9] and leaded by Euskaltzaindia, the Academy of the Basque Language. So, with the addition of new entries coming from the lexicographic databases of the above mentioned organizations, EDBL contains currently near 120,000 dictionary entries, more than 20,000 inflected forms (mostly verb finite forms) and about 700 non-independent morphemes, among others.

Concerning the information stored in the database, a full-fledged two-level morphology system for Basque is contained in it. Although EDBL contains mostly the general lexicon of standard Basque, lots of normative information are also stored in the database. EDBL contains many non-standard words and morphemes, such as dialectal forms, typical errors, etc., along with indications leading to their correct and standard use. The database is designed to gather other types of information of syntactic or semantic nature, such as subcategorization information, semantic features, etc., and it actually contains some of this kind of information in many cases.

Currently, EDBL resides under the ORACLE DBMS, on UNIX, and it may be consulted via the Internet (`http://ixa2.si.ehu.es/edbl`). Exportations from the database are made into XML documents, that are useful to create specifically formatted lexicons.

3.6 BasWN: Basque WordNet

The Basque WordNet is a lexical knowledge base that structures word meanings around lexical-semantic relations. It follows the specifications of WordNet [10],

[7] `http://www.uzei.com`
[8] `http://www.elhuyar.org/`
[9] `http://lexikoarenbehatokia.euskaltzaindia.net/cgi-bin/kontsulta.py`

as well as on its multilingual counterparts EuroWordNet and the Multilingual Central Repository (MCR). The Basque WordNet has been constructed with the expand approach [15], which means that the English synsets have been enriched with Basque variants. Besides, we also incorporate new synsets that exist for Basque but not for English. Due to EuroWordNet and the MCR frameworks, the Basque WordNet is already linked to the Spanish, Catalan, English and Italian wordnets, and it can also be linked to any other wordnet tightly linked to the English WordNet. The contents can be viewed using an interface which directly accesses the Basque, Spanish, Catalan and English WordNets[10]. It comprises 93.353 word senses and 59.948 words.

Up to now, the Basque WordNet has been focused on general vocabulary leaving aside specialized language and terminology. Nowadays we are creating a new resource called WNTERM (from WordNet and Terminology) with the aim of enriching the Basque WordNet with terminological information.

3.7 EPEC: Syntactically Annotated Text Corpus

EPEC Corpus (Reference Corpus for the Processing of Basque) is a 300,000 word corpus of standard written Basque [3] which aim is to be a training corpus for the development and improvement of several NLP tools [8]. EPEC has been manually tagged at different levels: morphosyntax, syntactic phrases, syntactic dependencies (BDT Basque Dependency Treebank) and WordNet word senses for nouns. In the course of the last few months, we started working to tag it with semantic roles.

The first version of this corpus (50,000 words) has already been used for the construction of some tools such as a morphological analyzer, a lemmatizer, or a shallow syntactic analyzer. This first version is publicly available in two websites:

- Ancora project[11]. This corpus can be downloaded and consulted with a friendly graphic interface.
- Natural Language Toolkit[12].

3.8 ZTC: Morphosyntactically Annotated Text Corpus

Today statistical tools for text processing are so powerful in language technology, that the number of words compiled and organized as text corpora could be used as a measure of the position of a language in the area.

The ZTC corpus [6] has been built by compiling text on the subject of *Science and Technology*. A previous inventory of years 1990-2002 registered 20 million words on this subject. The ZTC corpus compiled 10 millions words of standard written Basque. All those words were automatically annotated, and up to 1.8 million were manually revised and disambiguated. A specific interface for

[10] `http://ixa2.si.ehu.es/mcr/wei.html`
[11] `http://clic.ub.edu/ancora`
[12] `http://www.nltk.org`

advanced query of the corpus was also built. The result is a public resource: http://www.ZTcorpusa.net.

The creation of this resource would have been impossible without reusing the lemmatizer. We built a new tool for corpus compilation and annotation. The massive use of the lemmatizer was necessary.

The ZTC corpus is still far away from the size of the corpora for other languages; e.g., the BNC corpus[13], that is becoming a standard corpus resource, has 100 million words. However, the ZTC corpus is a very useful resource for manual study of Basque, as well as for machine learning techniques.

4 Conclusions

A language that seeks to survive in the modern information society requires language technology products. "Minority" languages have to do a great effort to face this challenge. Ixa group has been working since 1986 in adapting Basque to language technology, having developed several applications that are effective tools to promote the use of Basque. Now we are planning to define the BLARK for Basque [12].

From our experience we defend that research and development for less resourced languages should to be faced following this points: high standardization, open-source, reusing language foundations, tools, and applications, and incremental design and development of them.

We know that any HLT project related with a less privileged language should follow those guidelines, but from our experience we know that in most cases they do not. We think that if Basque is now in an good position in HLT is because those guidelines have been applied even though when it was easier to define "toy" resources and tools useful to get good short term academic results, but not reusable in future developments.

References

1. Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M., Oronoz, M., Sarasola, K., Soroa, A., Urizar, R.: A framework for the automatic processing of Basque. In: Proceedings of Workshop on Lexical Resources for Minority Languages (1998)
2. Aduriz, I., Alegria, I., Artola, X., Ezeiza, N., Sarasola, K., Urkia, M.: A spelling corrector for Basque based on morphology. Literary and Linguistic Computing 12(1), 31–38 (1997)
3. Aduriz, I., Aranzabe, M., Arriola, J.M., Atutxa, A., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., Urizar, R.: Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing. In: Archer, D., Rayson, P., Wilson, A., McEnery, T. (eds.) Proceedings of the Corpus Linguistics 2003 Conference, March 28-31, vol. 16 (1), pp. 10–11. Lancaster University, UK (2003)

[13] http://www.natcorp.ox.ac.uk

4. Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K.: Transfer-based MT from spanish into basque: Reusability, standardization and open source. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 374–384. Springer, Heidelberg (2007)
5. Amorrortu, E.: Bilingual Education in the Basque Country: Achievements and Challenges after Four Decades of Acquisition Planning. Journal of Iberian and Latin American Literary and Cultural Studies 2(2) (2002)
6. Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Sologaistoa, A.: ZT Corpus: Annotation and tools for Basque corpora. In: Corpus Linguistics, Birmingham (2007)
7. Artola, X., Díaz de Ilarraza, A., Soroa, A., Sologaistoa, A.: Dealing with Complex Linguistic Annotations within a Language Processing Framework. IEEE Transactions on Audio, Speech, and Language Processing 17(5), 904–915 (2009)
8. Bengoetxea, K., Gojenola, K.: Desarrollo de un analizador sintctico estadstico basado en dependencias para el euskera. Procesamiento del Lenguaje Natural 1(39), 5–12 (2007)
9. Borin, L.: Linguistic diversity in the information society. In: SALTMIL 2009 Workshop: IR-IE-LRL Information Retrieval and Information Extraction for Less Resourced Languages. University of the Basque Country (2009)
10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
11. Koskenniemi, K.: Two-level morphology: a general computational model for word-form recognition and production, University of Helsinki (1983)
12. Krauwer, S.: The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In: International Workshop Speech and Computer, Moscow, Russia (2003)
13. Petek, B.: Funding for research into human language technologies for less prevalent languages. In: Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece (2000)
14. Streiter, O., Scannell, K., Stuflesser, M.: Implementing nlp projects for noncentral languages: instructions for funding bodies, strategies for developers. Machine Translation 20(4), 267–289 (2006)
15. Vossen, P. (ed.): EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, Norwell (1998)
16. Williams, B., Sarasola, K., ÓCróinin, D., Petek, B.: Speech and Language Technology for Minority Languages. In: Proceedings of Eurospeech 2001 (2001)