A framework for the automatic processing of Basque

Aldezabal I., Ansa O., Arriola J.M.,
Díaz de Ilarraza A., Ezeiza N., Maritxalar A.,
Oronoz M., Sarasola K.,
Informatika Fakultatea
Euskal Herriko Unibertsitatea
649 PK–20080 DONOSTIA
e-mail.: jipsagak@si.ehu.es
Error! Bookmark not defined.

Aduriz I., Urkia, M.
UZEI
Aladapeta,20
20009 DONOSTIA
Tel: 34-43-473377
Error! Bookmark not defined.

1 Introduction.

In this paper we present the main research lines, tools and resources developed for the treatment of Basque by the IXA Group composed by 18 members from the Computer Science Faculty of the University of the Basque Country and by 2 members of UZEI (Basque Centre for Terminology and Lexicography) created in 1977 with the aim of promoting Basque lexicon's modernization within the normalization process of this language. Besides, the strategy followed by the group is presented.

Basque is a minority language. There are 700,000 Basque speakers They are around the 25% of the total population of Basque Country, but they are not regularly distributed. Most of them are bilingual nowadays. The four main literary dialects are: Guipuzcoan, Biscayan, Labourdin and Souletin. The Academy of the Basque Language (Euskaltzaindia) gave the first rules towards the standard Basque in 1968.

As Basque is a language with a very rich morphology[1], when we started working on the processing of Basque ten years ago we decided not to begin trying to develop advanced applications such as machine translation or query interfaces in natural language, we decided to develop broad basis dealing extensively with lexicon and morphology. Now those foundations became the basis for present and future developments. We distinguish four main sets among the current works in IXA group: In the first set, **applications**, we include those commercial systems oriented to not specialized users; in the second set, **tools,** we consider those systems which are NLP developers-oriented; the third group includes the **foundations**; and finally, the four set groups the systems and studies placed at a research level**: research.** Figure 1 shows this four groups and our

[1] Being Basque an agglutinative language, for the formation of words the dictionary entry independently takes each of the elements necessary for the different functions (syntactic case included). More specifically, the affixes corresponding to the determinant, number and case are taken in this order and independently of each other.

developments in each one of it are described in the next sections.

Sections 2, 3 and 4 describe respectively the foundations, tools and applications created by IXA group. Section 5 shows other research lines currently in development. Finally the paper ends with some concluding remarks.

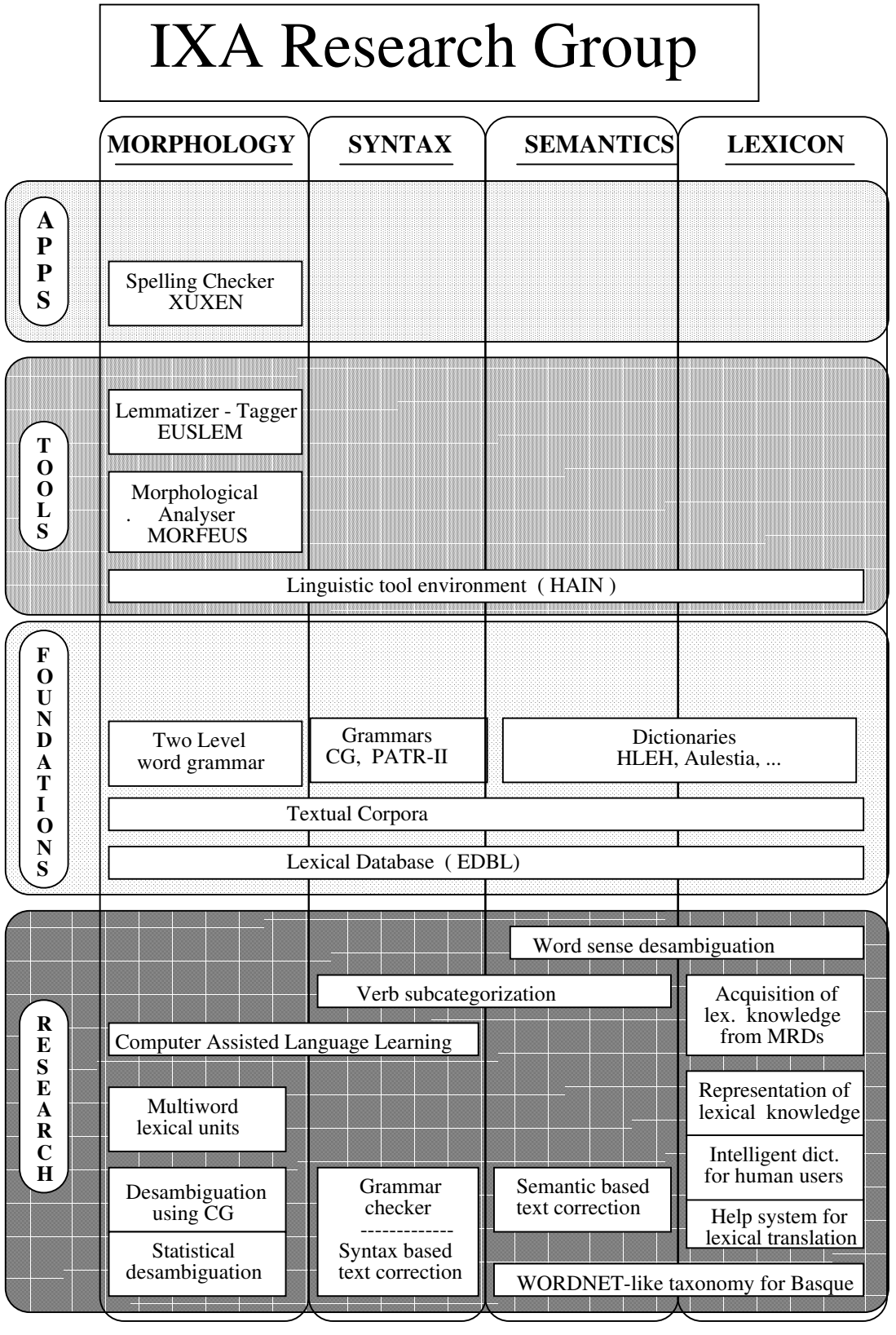

Fig.1.Products and systems sorted by application level and kind of knowledge.

1. Foundations

2.1 The lexical data-base EDBL

EDBL is a lexical database (LDB) for Basque (Agirre et al. 95). EDBL (Euskararen Datu-Base Lexikala) was created because of the peremptory need for sound lexical support for the construction of a general morphological analyzer and its most important by-product so far, the spelling checker/corrector Xuxen. The maintenance of the large amount of lexical information needed in such a project would not have been possible without a database management system. At present it contains over 70,000 entries, each with its associated linguistic features (category, subcategory, case, number, etc.). Other projects our group is currently involved in, like the construction of a lemmatizer/tagger—also derived from the morphological analyzer— or the development of a general syntactic parser require different types of lexical information. In this context, EDBL has been redesigned as a general basis for the multiple lexical needs that current and further work on the automatic treatment of Basque will have (we are involved so far in automatic processing tasks of written Basque).

The morphological usage of EDBL is not longer its central element, but one more of its several purposes. The main key of every item in the database is now composed by the headword and an homograph identifier, as in any conventional dictionary. The information is distributed in different parts, according to the different purposes it is intended to be used for.

The implementation of the database in a commercial RDBMS.

EDBL's main features are the following: 1) Multi-purpose. As said before, it has been designed as a general support and source for different applications and not for a unique application. In the short term, an explanatory dictionary containing definitions and examples (Sarasola I., 95) will be integrated into EDBL, thus enhancing it and converting it into an electronic dictionary suitable for human consultation. 2) Neutral. The linguistic descriptions held in it should not constrain any applications in the future. 3) Open and flexible. It will obviously allow the addition of new information; moreover, as its structure is based on feature structures. 4) User friendly. The interface for human users has to be designed as an easy-to-use tool. This interface is currently being developed.

2.2 Two level word grammar

The two-level model of morphology (Koskenniemi, 83) has become the most popular formalism for highly inflected and agglutinative languages. The two-level system is based on two main components:

- a lexicon where the morphemes (lemmas and affixes) and the possible links among them (morphotactics) are defined.
- a set of rules which controls the mapping between the lexical level and the surface level due to the morphophonological transformations

The rules are compiled into transducers, so it is possible to apply the system for both analysis and generation. We did our own implementation of the two-level model with slights variations, and applied it to Basque

(Alegria et al. 96a, b), a highly inflected and agglutinative language. Our two-level system has de following components: 1) 21 rules describing changes due to phonological, morphological, morphonological and orthographical reasons. 2) 44,384 items in the lexicon: 35,421 lemmas, 8,825 verb forms and 138 affixes, and 3) Those items are grouped into 135 sublexicons and with 99 continuation classes defined.

2.3 Grammars

In the way of computational treatment of a language, syntactic analysis is the task that necessarily follows the morphological one. Therefor, our group has already undertaken the syntactic challenge After an study of different formalisms for syntactic analysis, we chose to follow two different formalisms. We experiment with both formalisms and we think they can be seen as complementary methods.

2.3.1. PATR-II like unification grammar.

Unification based formalisms provide a linguistically and mathematically well-founded apparatus, tied with direct computational implementations. As there have not been many studies on unification grammars for Basque, we adopted the PATR-II formalism due to its flexibility for our proposes, giving the possibility of solving in a declarative way many linguistic problems that will have to be dealt with in other, more restrictive formalisms like LFG or HPSG. A first prototype grammar was implemented allowing wide coverage in the parsing of noun phrases. We plan to use this grammar in the detection of noun phrases, teaching syntax to Basque students, and in the implementation of a grammar checker. In order to treat efficiently grammar errors the system uses the gradual relaxation of syntactic constraints (Gojenola K. and Sarasola K., 94)

2.3.2. Constraint Grammar

The aim of Constraint Grammar (CG) is not, as it often happens, to create a toy grammar to play with laboratory sentences. CG is thought to analyze real text, that is, to be used with row text (Karlsson et al., 95). That philosophy agrees thoroughly with our aims, since the parser we are developing for Basque (Aduriz et al., 97b) must be the basis for other applications.

On the other hand, CG supports a parsing based on morphology. It goes without saying that this fits perfectly to a language like Basque where morphology and syntax are so related. A very important part of the analysis through CG is morphological disambiguation, namely the treatment of ambiguous output from morphological analysis using constraints based on linguistic knowledge.

We considered that CG formalism was appropriate to be used in general syntactic treatment. Far from the stiffness of other formalisms and, in spite of the problems, we think it is fit for languages with quite a free word order.

2.4 Corpus resources.

There are two main electronic corpora in Basque: Both of them are in UZEI to prepare dossiers about the origin and present use of the lexicon. Those dossiers are used by the Academy of the Basque Language

(Euskaltzaindia) to decide the unified lexicon. Both corpora is in a relational data-base (Oracle), but they are going to work in Basis in a few months to be offered via Internet..

1. The Corpus of the General Dictionary (OEH - Orotariko Euskal Hiztegia).

It is the historical corpus, the base for the general dictionary of the Academy. There are 310 books -complete works- with a total of 5.800.000 words. These books represent the Basque language from the 16th century to the middle of the 20th and are exhaustively copied out. Apart from the author and the title, the information about the period, dialect and text type is marked. It is not completely lemmatized, only when the Unified Lexicon Commission of the Academy asks for the dossiers about the history and variants of a word. But it can be consulted and it is possible to export data.

2. The Systematic Compilation of the Modern Basque (EEBS - Egungo Euskararen Bilketa-lan Sistematikoa)

EEBS contains 4.000.000 of lemmatized words showing the Basque lexicon of the 20th century (1900-1994). It is an statistical corpus: there are not full books, but randomly selected pages from near 5.000 works, with the aim of compiling as much as possible different words from multiple books and, inside them, from different chapters. All the documents (small parts of a work) contain the information about the author, title, (in the case of signed articles, the magazine -including the name, year, number and pages- where it was published), period, dialect, text type and size.

The origin of this sample is the classified (marked with the period, dialect, text type and size) inventory of all the publications in Basque language made in UZEI, from where it was extracted a cross-section sample, compiling works –proportionally- from all the periods, dialects and text types. Apart from that, there is another non-classified section with voluminous and fast publications, as newspapers, magazines, etc. These are not, often, written in a very "elaborated" language and there is not always the name of the author. This is an open corpus, it means, it is updated every year, because it's reason is to represent today's lexicon, to answer to the question: "What is used today...?".

The full corpus is lemmatized (with the standard lemma and the lemma of the variant), including complex words as compounds, idioms, etc. From 1900 to 1990 the corpus is marked in a homemade format, which is going to be passed (semi)automatically to SGML format. The last four years are coded in SGML format and lemmatized automatically using EUSLEM (The automatic lemmatizer for Basque).

3. Tools

The tools described in this section are not designed for common users, the potential users are researchers or others developers of linguistic applications.

3.1 Morphological Analyzer (Morpheus)

MORFEUS has a basic task in the automatic processing of Basque (Alegria et al. 96b). It assigns to each token in a text its lemma as well as all its possible morphological analysis. The rest of the modules will make use of that output so as to accomplish disambiguation and identify lexical units. This analyzer is based on the two-

level formalism.

First MORFEUS tries to analyze the word as **standard**. If the standard analyzer fails then the second module is activated and the analyzer tries to analyze it as a **variant** or **typical error**. If that module fails too, the third one analyses the word **without lexicon,** so that all the words have at least one interpretation.

This analyzer is a basic tool for current and future work on automatic processing of Basque and its first two applications are a commercial spelling corrector and a general purpose lemmatizer/ tagger.

The lexical transducers are generated as a result of compiling the lexicon and a cascade of two-level rules (Karttunen et al. 94). Their main advantages are speed and expressive power. Using lexical transducers for our analyzer we have improved both the speed and the description of the different components of the morphological system. Some slight limitations have been found too.

The number of interpretations given by both variant and lexicon-free modules is usually bigger than the one given in standard analyses. Therefore, and in order to avoid complicating further processes, a new module (the local disambiguation) discards some interpretations without taking into account the context.

Finally, the morphosyntactic treatment is applied, so that from the components of the word-form only the features actually necessary in syntax be raised. The output is given in text-format but we are currently working so as to give it in SGML format.

3.2 Lemmatizer (Euslem)

A lemmatizer-tagger is a computational tool used for assigning the correct lemma and grammatical category to each token of a corpus. It is a basic device for corpus analysis, automatic indexation, syntactic and semantic analyses etc. For example, the lemmatizer-tagger for Basque (EUSLEM) (Aduriz *et al.*, 94; Aduriz *et al.*, 96a) is essential for the Systematic Compilation of Modern Basque[2] (EEBS) project (Urkia, 97) being carried out by UZEI in the context of standardization of Basque.

The knowledge in computational linguistics of the former added to the large experience in lexicography of the latter is turning out to be vital for the completion of this project. Some previous outcomes resulted from their collaboration can be considered both the basis and antecedent of this work:

A tagset system has been developed. The one we have chosen for Basque is a three level system which the user can parameterize when using the program. In the first level seventeen general categories are included (noun, adjective, verb, etc). In the second one each category tag is further refined by subcategory tags. The last level includes other interesting morphological information (case, number, etc.).

If the lexicon-based analyzer produced no valid analysis, we needed a lexicon-free lemmatiser that wouldn't let any token unlemmatized or untagged, so that the system might be robust enough. Among the different

---

[2] During the 1987-1992 period UZEI manually compiled and lemmatised a three million word corpus of twentieth century's Basque texts, which is annually renewed.

systems we studied, we chose a two-level mechanism based on the idea used in speech synthesis (Black *et al.*, 1991).

In order to elaborate the lemmatiser/tagger for Basque (Aduriz et al., 96a, b), we have used the following components:

- A pre-processor to detect and tag figures, punctuation marks, etc. Pre-processing those elements is very useful because they don't produce ambiguous tags and, therefore, they reduce the strings of ambiguous elements.
- The general-purpose morphological analyzer for Basque, based on two-level morphology (Agirre *et al.*, 1992).
- Lexicon-free lemmatization so that the system may be robust.
- Treatment of compound lexical units.
- Disambiguation based on linguistic knowledge, completed by disambiguation based on statistics.

3.3 Linguistic tool Environment (Hain)

A prototype of a linguistic environment has been developed. The philosophy of the system is to make accessible simultaneously the linguistic tools described. This environment provides a graphical user interface that allows the user, among other operations, to look up in a dictionary, to analyze morphologically a word-form, to check if it is correct and to look for proposals. HAIN's main features are the following:

Multi-purpose. It has been designed as a general environment to integrate different applications and as a workbench for different users (linguists, Basque students, journalists, and so on).

Open and flexible, so that it will permit, at anytime, the integration of new tools.

User friendly and adaptive. Conceived for both programs and human users (specialized or not), the different interfaces corresponding to the tools have been designed as an easy-to-use environment.

This environment has been developed in a SUN Machine and the version for being accessible from the World Wide Web will be complete for next summer.

4. Applications

4.1 Spelling Checker/Corrector (Xuxen)

Xuxen is a spelling checker/ corrector for Basque (Agirre *et al.*, 92; Aduriz *et al.*, 93; Aduriz et al., 1997) based on two-level morphology). As mentioned, because Basque is an agglutinating and highly inflected language, a system based on a word-form list is not suitable.

The checker recognizes a word-form if a correct morphological breakdown is allowed.

- The treatment of orthographic errors is based on the parallel use of a two-level subsystem designed to detect previously typified misspellings, which is added to the two-level system used

by the morphological checker.

- The treatment of typographical errors is quite conventional; the proposals are fed into the spelling checker to check whether they are valid or not.
- Two-level morphology (Koskenniemi 83) has been the base for the design of the checker/ corrector. Only a slight adaptation was needed for the treatment of orthographic error and to improve the checking of proposals.
- The program has been developed in C and runs on a SUN SPARC machine, Macintosh and PC. It was commercialized in 1994.

  A filter program, buffers for the most frequent words (to improve the performance) and maintenance of the user's own dictionary are the essential elements added to the morphological system.
- In order to manage with precision the inclusion of new lemmas in the user's own dictionary, an interface for lexical knowledge acquisition has been designed.

The main existing correction method (to define a measurement for the distance between words and to calculate which words of the dictionary have a lesser distance) is not suitable for a lexicon system not based on a word-form list.

The treatment of typographical errors is quite conventional and performs the following steps:

- Generation of proposals to typographical errors using Damerau's classification (single deletion, insertion, substitution or transposition).
- Trigram analysis: proposals with trigrams below a certain probability threshold are discarded, while the rest are classified in order of trigrammic probability.
- Spelling checking of proposals; because on the basis of the previous criteria only, incorrect word-forms could be offered to the user.

5. Research and future work

Finally, we present other research lines currently developed in our group.

5.1 IDAZKIDE: A help system for the study of Basque

IDAZKIDE is an intelligent system that helps learners of Basque write their texts. This help is not only given at word level but also in some features sentence level. The system is based on a student model that gathers some futures of the students' learning process such as language level, student's favorite learning strategy, mother tongue, etc. All this information is extracted from a database completed by the teacher.

An intelligent system for analyzing a second language learning process is also being developed. That system is composed of two subsystems: the Knowledge Acquisition Subsystem and the Learning Process Subsystem. The database fits into the first subsystem and IDAZKIDE, the tool assisting in writing texts, fits into the

second one.

In the student model of IDAZKIDE we are also gathering information on the students' knowledge about a second language. We call that knowledge interlanguage (Díaz de Ilarraza et al., 97). To complete the interlanguage, we work with a corpus collected in some language schools of Basque (Maritxalar et al., 96). For the study of the corpus, we are adapting some linguistic tools developed by the IXA group such as the morphologic analyzer, and the lemmatiser (Maritxalar et al., 97).

5.2 IDHS (Intelligent Dictionary Help System)

IDHS (Intelligent Dictionary Help System) is a monolingual (explanatory) dictionary system (Artola & Evrard, 92). Its design was conceived from the study of questions that human users would like to be answered when consulting a dictionary. The fact that it is intended for people instead of automatic processing distinguishes it from other systems dealing with the acquisition of semantic knowledge from conventional dictionaries. The system provides various access possibilities to the data, allowing to deduce implicit knowledge from the explicit dictionary information. IDHS deals with reasoning mechanisms analogous to those used by humans when they consult a dictionary.

The starting point of IDHS is a Dictionary Database (DDB) built from an ordinary French dictionary. Meaning definitions have been analyzed using linguistic information from the DDB itself and interpreted to be structured as a Dictionary Knowledge Base (DKB). As a result of the parsing, different lexical-semantic relations between word senses are established by means of semantic rules (attached to the patterns); these rules are used for the initial construction of the DKB.

Once the acquisition process has been performed and the DKB built, some enrichment processes have been executed on the DKB in order to enhance its knowledge about the words in the language. Besides, the dynamic exploitation of this knowledge is made possible by means of specially conceived deduction mechanisms. Both the enrichment processes and the dynamic deduction mechanisms are based on the exploitation of the properties of the lexical semantic relations represented in the DKB (Agirre *et al.*, 94).

5.3 ANHITZ: A Multilingual Dictionary-System to Assist in the Translation

This doctoral thesis shows the most relevant aspects of an innovator multilingual dictionary-system named ANHITZ(Arregi et al. 94b). The system has been designed to help in the human translation, namely in the word-level translation. ANHITZ can offer both passive and active help, so that far from a single query-system it can be seen as an intelligent help-system.

This work brings together three different areas: the machine assisted human translation, the computational lexicography and the knowledge-based help-systems. From our view, it is necessary to integrate these areas, given that dictionaries should be specialized, active, adaptable and easy-to-use.

We have adapted the KADS methodology to model the use of dictionaries in the lexical translation. Based on KADS, four levels of knowledge —domain, inferences, tasks and strategy— have been distinguished in our model.

A way of representing the multilingual lexical-semantic knowledge has been proposed and, according to this representation, knowledge from dictionaries —two monolingual ones and one bilingual— has been stored in the domain-level. This proposal of representation is highly consistent and it could be used for whatever languages.

In the inference level the basic "behavior" of the dictionary knowledge is treated by means of lexical reasoning mechanisms, such as enrichments, lexical rules for dynamic deductions and operational basic functions.

Task descriptions of dictionary queries and their hierarchical organization have been managed in the strategic level. The conceptual model is the base for the design model and the system's architecture, in which the implementation aspects are defined. This architecture has adopted features of the active help systems in order to configure the external behavior of ANHITZ and its interaction with the human translator.

5.4 ITEM: Recovery and Extraction of Textual Information in a Multilingual Environment with Natural Language.

The main aim of this project is to explore and evaluate to which extent the use of NLP techniques can improve the processes of information recovery and extraction in textual or multimedia systems, enabling a multilingual access to them.

To develop these techniques we need resources and tools that allow to extract from any text the linguistic information necessary for the search, indexation and control processes that take part in information recovery. These resources and tools include a lexical database, a conceptual knowledge base, robust analyzers that permit to carry out efficiently, morphological analysis, tagging, syntactic and semantic analysis of texts (total or partial). As to multilingual access (Basque, Catalan, English, Spanish), we will focus on making possible that the search in a language allows to accede to the information in another one. Therefore, we will build a conceptual knowledge base that will connect the lexical databases of the different languages.

This project is being carried out in collaboration with the Natural Language Research Group from the Polytechnic University of Catalonia (UPC), the Computational Linguistics Group from Central University of Barcelona (UB) and UNED.

6. Conclusions

In this paper we have presented the different studies and products developed by our group during the last years. We think that these are important steps towards the process of making the language able to suit the modern life needs.

Once some basic tools have been developed and during the production of new ones, our efforts are concentrated now on providing access to the tools to all type of users: linguists, literate users, Basque students, translators and so on.

Until now a number of tools have been developed but important basis have been established for the

development of new ones.

Acknowledgements

These works have been supported from the Department of Economy of the Government of Gipuzkoa, the Department of Education of the Basque Government, the Commission of Science and Technology of the Spanish Government.

References

Aduriz, I., Agirre, E., Alegria, I., Arregi, X., Arriola, J.M., Artola X., Díaz de Ilarraza, A., Ezeiza, N., Maritxalar, M., Sarasola, K., Urkia, M. "Morphological Analysis Based Method for Spelling Correction" (poster session). Proceedings of the European association for Computational Linguistics. EACL 93. Utretch (Holland), 1993.

Aduriz, I., Alegria, I., Arriola, J.M., Artola X., Díaz de Ilarraza, A., Ezeiza, N., Urkia, M. "EUSLEM: un lematizador/etiquetador de textos en euskera". Boletín de la SEPLN, Córdoba, 1994.

Aduriz I., Alegria I., Arriola J. M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. "A corpus based morphological disambiguation tool for Basque". Boletín de la SEPLN, Revista no. 19, Sevilla, 1996a.

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. "EUSLEM: A Lemmatiser". Tagger for Basque" EURALEX'96, Göteborg (Sweden), 1996b.

Aduriz I., Aldezabal J.M., Artola X., Ezeiza N.,Urizar R. "MultiWord Lexical Units in EUSLEM, a lemmatiser-tagger for Basque". Papers in Computational Lexicography, Complex'96, Linguistics Institute, Hungarian Academy of Sciences, Budapest, 1996c.

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. "A spelling corrector for Basque based on morphology". Literary & Linguistic Computing, Vol. 12, No. 1, Oxford University Press, Oxford, 1997a.

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. "Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism". Recent Advances in NLP (RANLP97), Tzigov Chark (Bulgary), 1997b.

Agirre E., Alegria I., Arregi, X., Artola X., Díaz de Ilarraza A., Maritxalar, M., Sarasola K., Urkia M. "Xuxen: 1 Spelling Checker/Corrector for Basque based in Two-Level Morphology". Proceedings of ANLP'92, 119-125, Povo Trento, 1992.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola, K. "Lexical Knowledge Representation in an Intelligent Dictionary Help System". Proceedings of COLING'94, Kyoto (Japan), 1994a.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola, K. "Analysing world-level translation activity to design a computerised dictionary". Proceedings of Euralex'94, Amsterdam, 1994b.

Agirre E., Arregi X., Arriola J.M., Artola X., Díaz De Ilarraza A., Insausti J.M., Sarasola K."Different issues in the design of a general-purpose Lexical Database for Basque". First Workshop on Applications of Natural Language to Databases Versailles. 1995.

Alegria I., Artola X., Sarasola K., Urkia M. "Automatic morphological analysis of Basque". Literary & Linguistic Computing Vol. 11, No. 4, Oxford University Press. Oxford, 1996a.

Alegria I., Artola X., Ezeiza N., Gojenola K., Sarasola K. "A trade-off between robustness and overgeneration in morphology". Le traitement automatique du langage et les applications industrielles/Natural Language Processing and Industrial Applications. Volume I. pp 6-10. Moncton, Canada. 1996b.

Arregi X., Artola X., Díaz de Ilarraza A., Evrard F., Sarasola K. "Sistema Diccionarial Multilingüe: aproximación funcional". Boletín de la SEPLN, Santiago de Compostela, 1993.

Artola X., Evard, F. "Dictionnaire intelligent d'aide à la compréhension". Actas IV Congreso International EURALEX'90 (Benalmádena), 45-57, Barcelona, 1992.

Black A., Van de Plassche J., Williams B. "Analysis of unknown words through morphological decomposition". Proc. of the 5th Conference of the EACL. 1991

Díaz de Ilarraza A., Maritxalar M., Oronoz M. "An Implemented Interlanguage Model for Learners of Basque". Language Teaching and Language Technology Groningen, 1997. Forthcoming.

Gojenola K., Sarasola K. "Aplicación de la Relajación Gradual de Restricciones para la Detección y Corrección de Errores Sintácticos". Boletín de la SEPLN, Córdoba, 1994.

Karlsson F., Voutilainen A., Heikkila J. and Anttila A. "Constraint Grammar: A language independent system for parsing unrestricted text". Ed. Mouton de Gruyter 1995.

Karttunen L. "Constructing Lexical Transducers". Proc. of COLING'94, 406-411. 1994.

Koskenniemi K. "Two-level morphology: A general computational model for word-form recognition and production". PhD Thesis. Univerity of Helsinki. 1983.

Maritxalar M. Díaz de Ilarraza A., Alegria I., Ezeiza N. "Modelización de la competencia gramatical en la interlingua basada en el análisis de corpus". Boletín de la SEPLN, Revista no. 19, Sevilla, 1996.

Maritxalar M., Díaz de Ilarraza A., Oronoz M. "From Psycholinguistic Modelling of Interlanguage in Second Language. Acquisition to a Computational Model". Computational Natural Language Learning (CoNLL97). In conjunction with ACL'97/EACL'97 Madrid, Spain, 11-12th July 1997.

Urkia M. "Euskal morfologiaren tratamendu morfosintaktikorantz". PhD Thesis. Univerity of the Basque Country. 1997.

Dictionaries

Sarasola, I. "Hauta-Lanerako Euskal Hiztegia". GK. 1995.