Towards the definition of a basic toolkit for HLT

Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sarasola K., Soroa A

IXA Group
Dept. of Computer Languages and Systems
University of the Basque Country, 649 P. K.,
E-20080 Donostia, Basque Country

KSarasola@si.ehu.es

Abstract

This paper intends to be an initial proposal to promote research and development in language independent tools. The definition of a basic HLT toolkit is vital to allow the development of lesser-used languages. Which kind of public HLT products could be integrated, at the moment, in a basic toolkit portable for any language? We try to answer this question by examining the fifty items registered in the Natural Language Software Registry as language independent tools. We propose a toolkit having standard representation of data and develop a strategy for the integration, in a common framework, of the NLP tools.

1. Introduction

SALTMIL, the ISCA SIG (International Speech Communication Association Special Interest Group) on Speech and Language Technology for Minority Languages, has the overall aim of promoting research and development in the field of speech and language technology for lesser-used languages. Actually, its main activity is providing a channel of communication between researchers by means of workshops and the discussion list. The members of SALTMIL, we often wonder how to promote research and development in a more active way. In this paper we would like to propose a medium term project to accomplish that goal: the definition of a basic toolkit for HLT. Of course, this toolkit should be designed following the basic principles of reusability and portability[1]. So, the adoption of common standards and procedures will help to minimise costs and workload in research. This way will be beneficial for any kind of language (and vital for lesser-used languages), and would define a new collaboration-space for researchers working with different languages.

The real challenge is, however, how to define a basic toolkit for HLT? In this paper we will not resolve this problem, but we want to lay some foundations to address it. First, we will try to collect an initial list of present tools and applications that are portable (usable) for different languages:

- How many of the present HLT tools and applications are portable?
- How many of them are free for academic and public uses?
- Is there any tool for any of main basic applications? or… Is there any application with no accessible tool?

In this way, by recognizing which are the most basic tools, we propose four phases as a general strategy to follow in the processing of any language. Therefore, tools considered in the first phase will be taken as more basic than the later ones.

The paper is organized as follows: Section 2 proposes a strategy to develop language technology for language, grouping linguistic resources, tools and applications in four different phases. Section 3 examines the programs registered by the Natural Language Software Registry (NLSR) in order to determine the present proportion between portable and not-portable HLT products. Section 4 proposes a standard representation of linguistic data; it is a method we use in IXA Group in order to allow the integration between different tools in the same HLT framework; the standard representation would be fundamental for any possible basic toolkit. Finally, some concluding remarks are included.

2. Recognizing basic tools and their preference

We present here an open proposal for making progress in Human Language Technology. This proposal is based on the fifteen years experience of the IXA Group with the automatic processing of Basque. Anyway, the steps here proposed do not correspond exactly with those observed in the history of the processing of English, it is due to the high capacity and computational power of present computers allows arranging problems in a different way. We must remark that our work has been centered on the processing of written language and that we do not have any reliable experience on spoken language. However, in this proposal some general steps on speech technology have included.

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 research and improving language foundations. Therefore, these three levels (language foundations, tools and applications) have

[1] Main themes chosen for the last two ISCA SALTMIL SIG workshops were "*Re-usability and strategic priorities*" (Athens 2000) and "*Portability Issues in Human Language Technologies*" (Gran Canaria 2002).

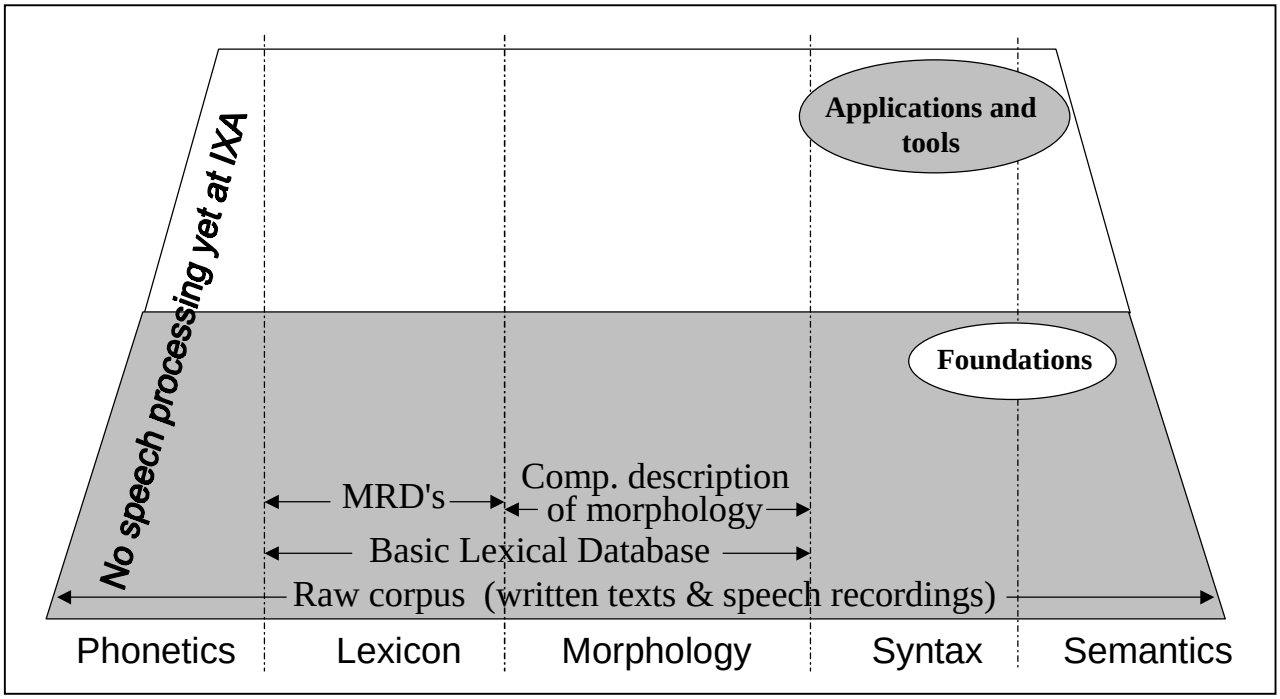


Figure 1. First phase: Foundations.

to be incrementally developed in a parallel and coordinated way in order to get the best benefit from them. Taking this into account, we propose four phases as a general strategy to follow in the processing of the language.

Initial phase: Foundations (see Figure 1).

- Corpus I. Collection of raw text without any tagging mark.
- Lexical database I. The first version could be simply a list of lemmas and affixes.
- Machine-readable dictionaries.
- Morphological description.
- Speech corpus I.
- Description of phonemes.

Second phase: Basic tools and applications.

- Statistical tools for the treatment of corpus.
- Morphological analyzer/generator.
- Lemmatizer/tagger.
- Spelling checker and corrector (although in morphologically simple languages a word list could be enough).
- Speech processing at word level.
- Corpus II. Word-forms are tagged with their part of speech and lemma.

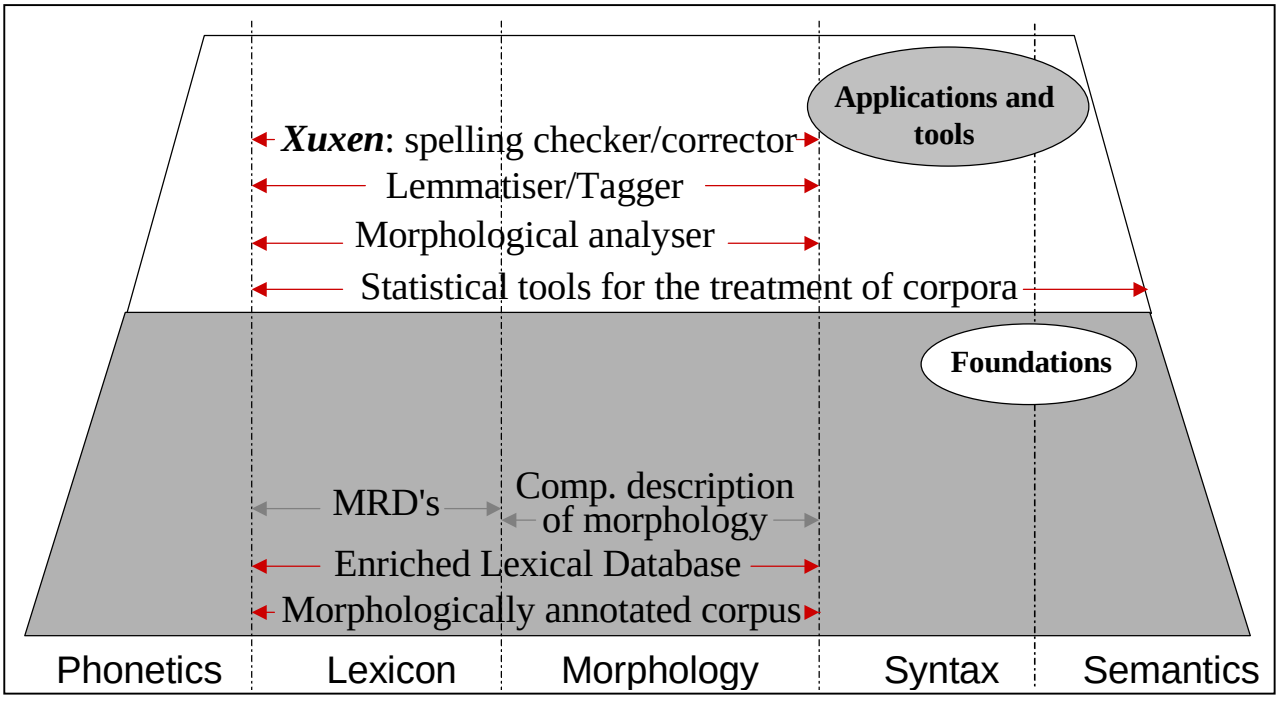


Figure 2. Second phase: Basic tools and application.

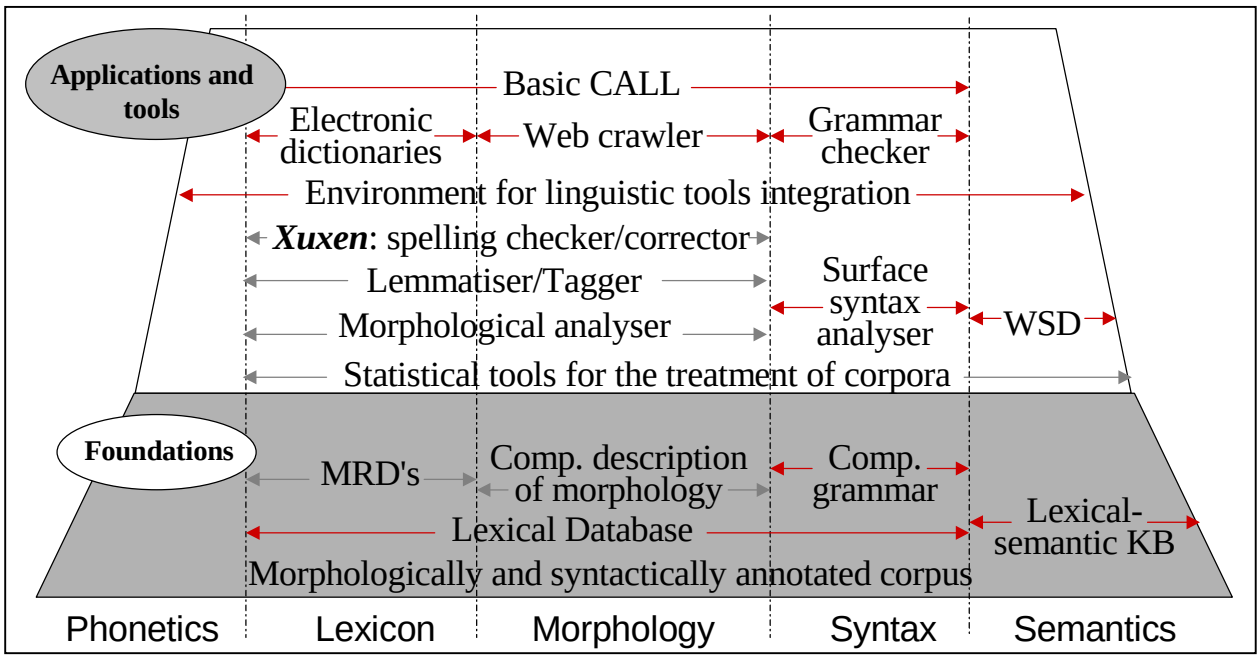


Figure 3. Third phase: advanced tools and applications.

- Lexical database II. Lexical support for the construction of general applications, including part of speech and morphological information.

Third phase: Advanced tools and applications.

- An environment for tool integration. For example, following the lines defined by TEI using XML. Section 4 describes this proposal.
- 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.
- 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.
- Basic Computer Aided Language Learning (CALL) systems.

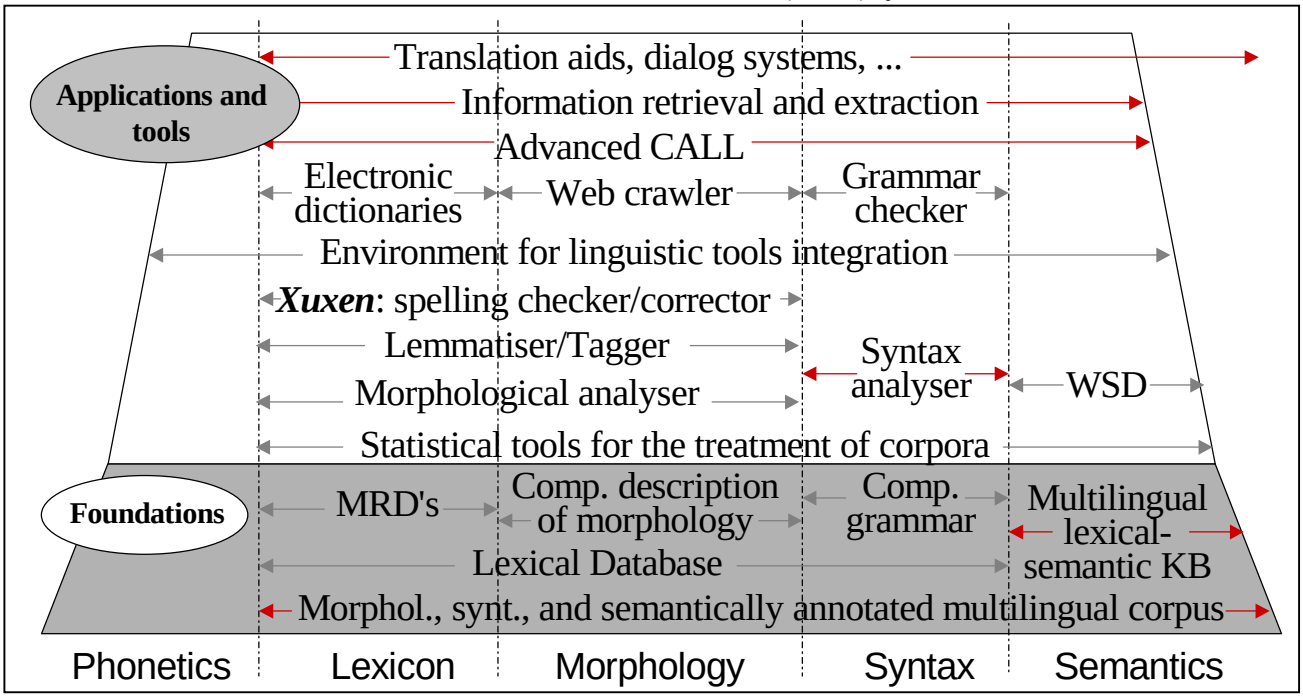
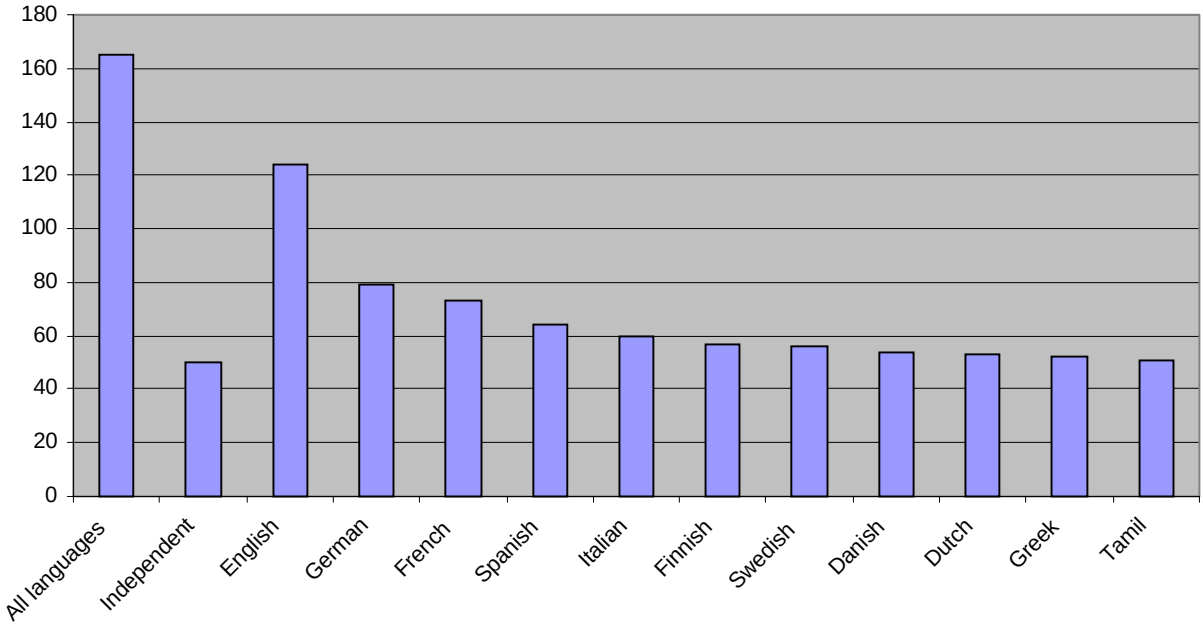


Figure 4. Fourth phase: Multilingualism and general applications..



Fourth phase: Multilingualism and general applications.

- Information retrieval and extraction.
- Translation aids. Integrated use of multiple on-line dictionaries, translation of noun phrases and simple sentences.
- Corpus IV. Semantically tagged text after word-sense disambiguation.
- Knowledge base on multilingual lexico-semantic relations and its applications.
- Dialog systems.

Now that we have started working on the fourth phase, every foundation, tool and application developed in the previous phases is of great importance to face new problems.

3. Present portable HLT products

Which is the start point at the present? Which kind of public HLT products could be integrated, at the moment, in a basic toolkit portable for any language?

With the aim of looking for data to answer to those questions, we examined the programs registered in the Natural Language Software Registry[2] (NLSR), an initiative of the Computational Linguistics Association (CL) and hosted at DFKI in Saarbrücken. The NLSR concentrates on listing HLT software, but it does not exclude the listing of linguistic resources (corpus, monolingual and multilingual lexicon). Other institutions, such as ELRA/ELDA or the Linguistic Data Consortium, provide listings of such resources. However, looking for portable products, to be precise, looking for products usable for multiple languages, the NLSR result sufficient because, actually, all linguistic resources are related to particular languages and so, they are not significant in this search. Of course, there are other HLT tools that have not been submitted to the NLSR, but we think that examine this database is a good start point.

3.1. Present proportion between portable and not-portable HLT products

First of all, we looked for how many of the present HLT tools and applications support different languages. This task was not very difficult because the system allows queries with a particular value for the slot named *Supported language(s)*. Figure 5 shows that a) the all amount of programs registered is 167; b) 50 of them (30%) has been declared to be language independent; c) of course, English is the language that support most of the programs. 125 support English (75%), that means that only 42 systems have been defined for the remaining 24 languages defined in NLRS; d) German, French, Spanish and Italian are the next languages an they are supported only by 79, 73, 64 and 60 respectively; and e) other languages are supported by those fifty defined as language independent and, occasionally, by a few other programs, for example 51 hits for Tamil. Those data reveals evident the significance of portability in Natural Language Software.

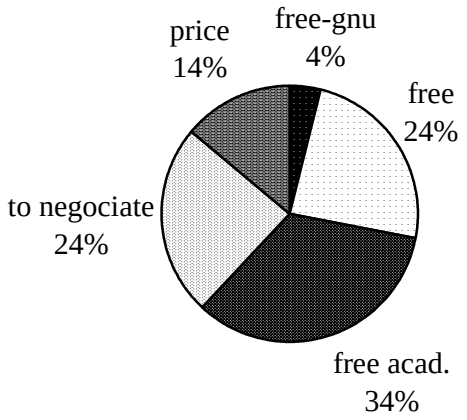


Figure 6: Price of portable HLT products

[2] http://registry.dfki.de

3.2. Price of portable HLT products

How many of the portable HLT products are free for academic and commercial uses? Among the fifty products they are 14 programs that free for any use (two of them, Zdatr and the speech synthesizer MBROLA, are distributed under the GNU Public Public License). Other 17 systems are free for academic uses. The price of 12 systems is defined as "to negotiate" even for academic uses. And finally 7 systems has a fixed price stated from $129 to $799; their average price is $546.

3.3. Distribution of portable products between HLT sections

Is there any portable tool for all the main basic sections in HLT? Or… is there any application with no accessible tools? Table 1 shows the distribution by sections of language independent software in NLSR. Similar data is shown for products that support English. We remark the following points: a) the number of products for the last four sections is not enough to be considered: b) the distribution of language independent products is similar to that of the total amount of products; c) there is any system in every section; d) the percentage of language independent products is considerable higher in Spoken Language and in NLP Development Aid.

Section	Total	Indep.	% indep.	Eng.	% Eng.
Total	167	50	0,30	125	0,75
Annotation	15	4	0,27	13	0,87
Written lang.	122	28	0,23	90	0,74
Spoken language	31	15	0,48	23	0,74
NLP development Aid	41	16	0,39	31	0,76
Lang. Resources	23	6	0,26	18	0,78
Multimedia	2	1	0,50	1	0,50
Multimodality	5	1	0,20	4	0,80
Evaluation	4	3	0,75	4	1,00

Table 1: Distribution of software by HLT sections

And now let's consider the distribution of NSLR products taking into account the kind of linguistic knowledge they manage. The kinds of knowledge to be considered are those referred in the previous section plus special points for NLP frameworks than includes facilities for lexical, morphology, syntax or speech. There is not any program to deal with dictionaries (creation of structured versions of dictionaries or integration of them in other applications), nor for semantics.

3.3.1. Corpus

Product	Description	Price
Alembic Workbench	a multi-lingual corpus annotation development tool	free
Bigram Statistics Package	Bigram analysis software	free
emdros	text database engine for linguistic analysis and research	free
PWA	Word Aligner	free acad.
SRILM -- SRI Language Modeling Toolkit	Statistical language modeling toolkit	free acad.
Entropizer 1.1	A toolbox for sequential analysis	to negotiate

Table 2: NLSR language independent products for corpus

3.3.2. Morphology

Product	Description	Price
PC-KIMMO	Two-level morphological analyzer	free acad.
TnT - Statistical Part-of-Speech Tagging	a statistical part-of-speech tagging for german, english and languages that delimit words with space	free acad.

Table 3: NLSR language independent product for morphology

3.3.3. Lexical databases

Product	Description	Price
DATR	A formalism for lexical knowledge representation	free
Xerox TermOnLine	Xerox TermOnLine is a terminology database sharing tool	to negotiate
Xerox TermOrganizer	Xerox TermOrganizer is a terminology database management system.	to negotiate

Table 4: NLSR language independent product for lexical databases

3.3.4. Speech

Product	Description	Price
IVANS: The Interactive Voice ANalysis System	Voice analysis, voice quality rating, voice/client data management	$749
CSRE - Computerized Speech Research Environment	speech analysis, editing, synthesis and processing system	$750
The OroNasal System	Nasalance measurement, analysis of oral and nasal airflow/energy in speech	$799
CSLU Toolkit	a comprehensive suite of tools to enable exploration, learning, and research into speech and human-computer interaction	free acad.
CSL -- Computerized Speech Lab	speech acquisition, analysis and playback	to negotiate
Signalyze(tm)	Interactive program for speech/signal analysis (runs only on Macintosh)	$350
TFR: The Time-Frequency Representation System	a comprehensive speech/signal analysis, editing and processing system	$599
Multi-Speech	a comprehensive speech recording, analysis, feedback, and measurement software program	to negotiate
WinPitch, WinPitch II	Speech analysis and annotation	to negotiate
ProTrain	speech analysis and speech production training system	$349
Praat	a research, publication, and productivity tool for phoneticians	free acad.
MBROLA	a speech synthesizer based on the concatenation of diphones	free-GNU
EULER	a freely available, easy-to-use, and easy-to-extend, generic multilingual TTS	to negotiate

Table 5: NLSR language independent product for speech

3.3.5. Syntax

Product	Description	Price
ASDParser and ASDEditor	Parser and editor for Augmented Syntax Diagram grammars, implemented in Java.	free
XLFG	Syntactic analysis using the LFG formalism	free
AGFL Grammar Work Lab	Formalism and tools for context free grammars	free acad.
CUF	constraint-based grammar formalism	free acad.
GULP -- Graph Unification Logic Programming	an extension of Prolog for unification-based grammar	free acad.
LexGram	development and processing of categorial grammars	free acad.

Table 6: NLSR language independent product for syntax

3.3.6. NLP framework

Product	Description	Price
Alembic	an end-to-end multi-lingual natural language processing system	free
The Quipu Grok Library	a library of Java components for performing many different NLP tasks	free
PAGE: A Platfrom for Advanced Grammar Engineering.	System for linguistic analysis and test of linguistic theory (HPSG, FUG, PATR-II). Can be used as part of a deep NLP system or as part of a speech system (a special version is used in Verbmobil).	to negotiate
TDL---Type Description Language	System for linguistic analysis and test of linguistic theory (HPSG, FUG, PATR-II). Can be used as part of a deep NLP system or as part of a speech system (a special version is used in Verbmobil).	to negotiate
QDATR	An implementation of the DATR formalism	free acad.
Kura	Kura is a system for the analysis and presentation of linguistic data such as interlinear texts.	free
Zdatr	Zdatr is a standardised DATR implementation in ANSI C	free-GNU

Table 7: NLSR language independent product for NLP frameworks

3.3.7. Applications

Product	Description	Price
BETSY - Bayesian Essay Test Scoring sYstem	Free Windows based text classifier/essay scorer	free acad.
Flag	Terminology, style and language checking	to negotiate
Universal Translator Deluxe	An omni-directional translation system	$129
Onix	High performance information retrieval engine	to negotiate
Brevity	Document summarization toolkit	to negotiate

Table 8: NLSR language independent product for applications

4. A standard representation for linguistic data using TEI conformant feature structures

The standard representation of linguistic data in order to allow the integration between different tools in the same HLT framework will be fundamental for any possible basic toolkit. In this section we present as a proposal the strategy used for the integration, in a common framework, of the NLP tools developed for Basque during the last twelve years (Artola et al.; 2000). The documents used as input and output of the different tools contain TEI-conformant feature structures (FS) coded in SGML[3]. These FSs describe the linguistic information that is exchanged among the integrated analysis tools.

The tools integrated until now are a lexical database, a tokenizer, a wide-coverage morphosyntactic analyzer, a general purpose tagger/lemmatizer, and a syntactic parser.

Due to the complexity of the information to be exchanged among the different tools, FSs are used to represent it. Feature structures are coded following the TEI's DTD for FSs, and Feature System Declaration (FSD) descriptions have been thoroughly defined.

The use of SGML for encoding the I/O streams flowing between programs forces us to formally describe the mark-up, and provides software to check that this mark-up holds invariantly in an annotated corpus.

A library of Abstract Data Types representing the objects needed for the communication between the tools has been designed and implemented. It offers the necessary operations to get the information from an SGML document containing FSs, and to produce the corresponding output according to a well-defined FSD.

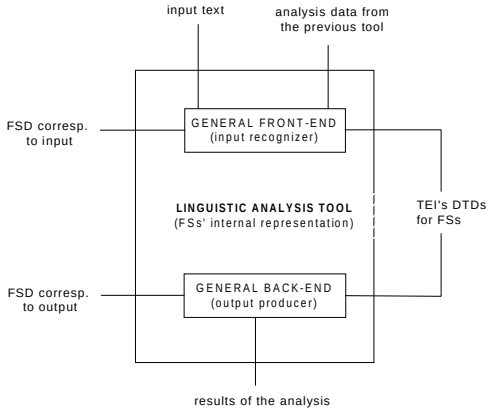


Figure 7. Schematic view of a linguistic analysis tool with its general front-end and back-end.

The use of SGML as an I/O stream format between programs has, in our opinion, the following advantages:

a) It is a well-defined standard for the representation of structured texts that provides a formal framework for the internal processing.
b) It provides widely recognized facilities for the exchange of data: given the DTD, it is easy to process any conformant document.
c) It forces us to formally define the input and the output of the tools used for the linguistic analysis of the text.
d) It facilitates the future integration of new tools into the analysis chain.
e) Pieces of software are available for checking the syntactic correctness of the documents, information

[3] All the references to SGML in this section could be replaced by references to XML.

retrieval, modifications, filtering, and so on. It makes it easy to generate the information in different formats (for processing, printing, screen-displaying, publishing in the web, or translating into other languages).

f) Finally, it allows us to store different analysis sets (segmentations, complete morphosyntactic analyses, lemmatization results, and so on) linked to a tokenized piece of text, in which any particular analysis FS will not have to be repeated.

.

5. Conclusions

If we want HLT to be of help for more than 6000 languages in the world, and not a new source of discrimination between them, the portability of HLT software is a crucial feature. Looking for language independent software in the Natural Software Registry, we saw that only 30% of the tools has been so declared; that 62% of those language independent programs are at least academic free and that they are quite homogeneously distributed among the different sections of HLT and among the kinds of knowledge they manage.

As many problems would arise when trying to coordinate several of those language independent programs, we present as a proposal the strategy used for the integration, in a common framework, of the NLP tools developed for Basque. Feature structures are used to represent linguistic information, and feature structures are coded following the TEI's DTD for FSs, and Feature System Declaration descriptions (FSD) have been thoroughly defined.

Worldwide international organizations that work for the development of culture and education should promote the definition and creation of a basic toolkit for HLT available for as many languages as possible. ISCA SALTMIL SIG should coordinate researchers and those organisations to initiate such project.

References

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Soroa A.. A Proposal for the Integration of NLP Tools using SGML-tagged documents. In *Proc. of the Second Int. Conf. on Language Resources and Evaluation*. Athens (Greece). 2000

Petek B. "Funding for research into human language technologies for less prevalent languages" Second International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece.

Sarasola K. "Strategic priorities for the development of language technology in minority languages". Proceedings of Workshop on "Developing language resources for minority languages: re-useability and strategic priorities". Second International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece.