Strategic priorities for the development of language technology in minority languages

K. Sarasola

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

jipsagak@si.ehu.es

Abstract

Language technology development for minority languages differs in several aspects from their development for widely used languages. The high capacity and computational power of present computers, added to the scarcity of human and linguistic resources implies the design of new and different strategies. This proposal presents the conclusions after twelve years of experience with the automatic processing of Basque.

1. Introduction

Language Engineering is recognized as one of the fundamental enabling technologies for the future. Language Engineering will make an indispensable contribution to the success of the information society. The availability and usability of new telematic services will depend on developments in Language Engineering. In the future natural language will become a standard computer interface providing us with the facility to communicate with a range of devices, included our computer, and to do so in our native language. But most of the working applications are available only in English. Minority languages have to do a great effort to face this challenge. In this paper we make an open proposal for making progress in Language Engineering. The steps here proposed do not correspond exactly with those observed in the history of the English processing, because the high capacity and computational power of present computers allows arranging problems in a different way. We define this strategy based on the experience of the IXA group with Basque.

Section 2 describes the phases we propose for the development of language technology. Section 3 suggests what not to do when working on the treatment of minority languages. Finally the paper ends with some concluding remarks.

2. Phases in the development of language technology

We distinguish three main levels among the works on Language Engineering. In the first level, **applications**, we include those commercial systems oriented to non-specialized users; in the second level, **tools,** we consider those systems that are oriented to application developers; and finally, the third group includes the **language foundations.** 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 have to be incrementally developed in a parallel and coordinated way in order to get the best benefit from them.**

Figure 1 shows these three levels (applications, tools and foundations), where each item is placed in different columns with respect to the linguistic knowledge it needs.
We propose five phases as a general strategy to follow in the processing of the language. Different phases are represented in Figure 1 by means of shade-levels: the items to be created in the first phase are presented in white boxes, while higher phases are distinguished by increasing gray percentages. The items are the following:

Initial phase: Laying foundations:

- Corpus I. Collection of raw text without any tagging mark.
- Lexical database I. It is the first version, which could be simply a list of lemmas and affixes.
- Machine-readable dictionaries. Bilingual and monolingual dictionaries, thesaurus, …
- Morphological description. Formalization of morphological phenomena. It is absolutely necessary for agglutinative languages.
- Speech corpus I. Collection of speech recordings.
- Description of phonemes.

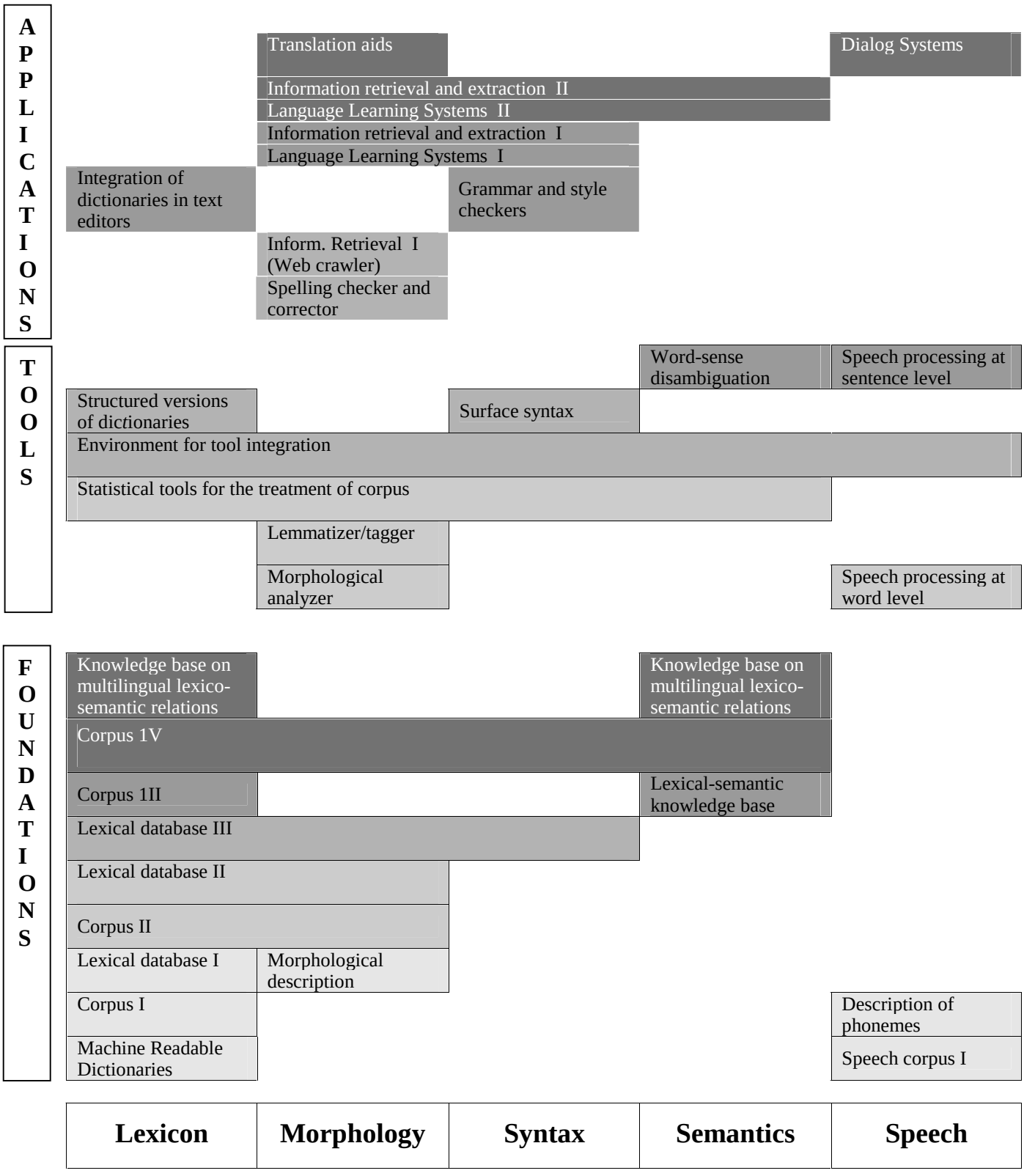

Figure 1: Phases in the development of language technology

Fifth phase: Multilinguality and general applications
Fourth phase: Advanced tools
Third phase: Tools of middle complexity
Second phase: Basic tools
Initial phase: Laying foundations

Second phase: Basic tools

- Statistical tools for the treatment of corpus: bigram and trigram frequencies, word count, collocations, co-occurrences, …
- Morphological analyzer/generator. It must be able to analyze or generate every word-form, giving the sequence of their morphemes.
- Lemmatizer/tagger. Based on the morphological analyzer it is able to disambiguate among different morphological readings for a word taking its context into account.
- Speech processing at word level.
- Corpus II. The word-forms in this second version are tagged with their corresponding part of speech and lemma.
- Lexical database II. Lexical support for the construction of general applications. This second version includes the part of speech, and morphological information (such as possible combination of morphemes, case, number, tense or aspect, …).

Third phase: Tools of medium complexity

- An environment for tool integration. It allows the integrated use of the available tools. An standard representation of linguistic knowledge is needed for the communication among tools. For example, following the lines defined by TEI (Text encoding Initiative) using SGML.
- Spelling checker and corrector. These will be developed using the lexical database and the morphological analyzer (although in morphologically simple languages a word list could be enough).
- Web crawler. Traditional search machine that integrates lemmatization and language identification and manages different formats: html, txt, doc, …
- Surface syntax. Recognition of simple syntactic constituents such as verbs, noun phrases or prepositional phrases.
- Structured versions of dictionaries. Databases allowing sophisticated queries. For example, asking for entries ending with "able" which are adjectives and contain the word "compare" in their definition.
- Lexical database III. The previous version is enriched with multiword lexical units.

Fourth phase: Advanced tools

- Corpus III. Syntactically tagged text.
- Grammar and style checkers.
- Integration of dictionaries in text editors.
- Lexical-semantic knowledge base. Creation of taxonomy of concepts. (e.g.: Wordnet)
- Word-sense disambiguation.
- Speech processing at sentence level.
- Language learning systems.

Fifth phase: Multilinguality and general applications

- Corpus IV. Semantically tagged text after word-senses have been disambiguated.
- Information retrieval and extraction.
- Translation aids. Integrated use of multiple on-line dictionaries, translation of noun phrases and simple sentences.
- Dialog systems.
- Knowledge base on multilingual lexico-semantic relations and its applications. The objective is to connect equivalent concepts or words from taxonomy representations in different languages.

There is a previous and necessary phase for languages that do not use Latin characters or even use a non-standard way to write words. In those cases previous work is needed to define the written representation of words.

3. What not to do

Do not start developing applications if linguistic foundations are not defined previously, that is, we recommend to follow the order given above: foundations, tools and applications.

When a new system must be planned do not create ad hoc lexical or syntactic resources. Design those resources in a way that they could be easily extended to full coverage and reusable by any other tool or application. For example, as Basque is a language with a very rich morphology, when we started working on the automatic processing of Basque we decided not to begin with advanced applications such as machine translation or natural language interfaces, but rather to develop a broad foundation based on the lexicon and morphology. Now those foundations have become the basis for present and future developments.

When 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. We know we will not become rich with those products, and market criteria do not usually apply, so that it is desirable to avoid needless and costly repetition of work.

4. Conclusion

This proposal presents the conclusions after twelve years of experience with the automatic processing of Basque. We present a long-term strategy arranged sequentially in five phases. 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 and applications.

5. References

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., ArtolaX., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. (1998). *A Framework for the Automatic Processing of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation, Granada (Spain).

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Soroa A. (2000). 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).

Baker P., McEnery T., Sebba M., and Lou Burnard. (1998). Minority Language Engineering. In ELRA Newsletter, Vol 3 N4 Nov 1998.

Calzolari, N. (1998). An overview of written Language Resources in Europe: a few reflections, facts and a vision. In *Proc. of the first Int. Conf. on Language Resources and Evaluation*. Granada (Spain).

Höge, H. (1998). Spoken languages resources for voice driven man machine interfaces. In *Proc. of the first Int. Conf. on Language Resources and Evaluation*. Granada (Spain).

Ostler, N. (1999) Does Size Matter? Language Technology and the Smaller Language. ELRA Newsletter, Vol 4 N1 Jan-Mar 1999. Paris: European Language Resources Association.

Somers, H. (1998)..Language Resources and Minority Languages. In Language Today. Number 5, 1998. Nottingham, UK: Language Publications Ltd. pp. 20-24.