Techno-langue: The French National Initiative for Human Language Technologies (HLT)

Stéphane Chaudiron & Joseph Mariani
Ministère délégué à la Recherche
1, rue Descartes – 75231 Paris cedex 05 - France
{firstname.lastname}@technologie.gouv.fr

Abstract

Techno-langue is the French National Program on HLT supported by the French ministries in charge of Research, Industry and Culture. It addresses four action lines: creating basic language and software resources, organizing evaluation campaigns, participating in the standardization process and creating a Web Portal for disseminating information and surveys to a large audience. This paper presents the main results of the program and an ongoing initiative for launching a transnational program at the European level on a similar basis.

1. Introduction

Following a report to the Prime Minister in 2000 concerning the major role of HLT in the Information Society, the *Techno-langue* program has been launched in April 2002 as a large French national program on Language Technologies (LT).

Techno-langue is very closely linked to related existing R&D programs in the field of Information and Communication Technologies, so-called the Technological Research & Innovation Networks (RRIT) in Telecommunications, Software Technologies, and Audiovisual & multimedia. Beside the RRIT in charge of financing R&D projects in a cooperative way between academics and industrials, the *Techno-langue* program focuses on building an infrastructure for creating basic language and software resources, organizing evaluation campaigns for HLT and stimulating participation in standardization committees. Another concern of the program is to help dissemination of HLT information through a dedicated Web Portal.

This paper presents the results of the program along these four action lines and we conclude by pointing out the need for a large European initiative, of the same nature.

2. Context and rationale

Techno-langue is a three-year program which will be fully completed in 2006. A steering committee of 15 members equally representing the NLP and speech communities, and the companies and academic laboratories, is in charge of supervising the review of the proposals and to steer the program.

The overall funding budget dedicated to the program is over 7.5 millions euros coming from the three ministries in charge of Research, Industry and Culture. The global effort, made both by the national bodies and the industrial sector, reaches about 11.5 M€ as all projects are funded on a share-cost basis.

Techno-langue is concerned by four action lines:

- The first one aims at stimulating the production and diffusion of Language Resources and basic language processing tools. It aims at the emergence of a "toolbox" containing the minimal linguistic and software resources necessary for the automation of the French language. One major aspect is the need for a wide distribution of these resources;
- Secondly, *Techno-langue* addresses the Evaluation topic by funding the organization of evaluation campaigns. More generally, the program aims at creating an infrastructure for evaluating HLT technologies.
- The third goal is to support the French participation in standardization committees at the international level and to disseminate the negotiation results to the actors.
- The last action line concerns the creation of a Web Portal on HLT in order to ensure a permanent technological, scientific and industrial watch by making available the results of the projects and by disseminating the news concerning the HLT field, both from the academic and industrial sectors.

Late 2002, 27 projects have been short-listed after the evaluation process from 52 proposals, and 21 projects have finally been funded. More than 90 different participants are involved in the projects: 33 from industry, 39 public research centers, 11 others (Associations, CEA, French DoD…) and 11 from outside France (Bell Labs (USA), NII (Japan), EPFL and LATL (Switzerland), RALI (Canada) …), which take part in the program on a self funding basis.

3. Results of the program

The 21 funded projects have been organized in five clusters. We present hereafter a summary of the results. It is possible to have access to a more complete description of the different projects on the Techno-langue website: www.technolangue.net

3.1. Creation of basic language and software resources

The first three clusters, called AGILE, NEOLOGOS and DICTIONNAIRE, are concerned by the creation and dissemination of basic resources. Within the **AGILE** cluster, four projects have been funded:

3.1.1. TILT

Project leader: ATILF and AFNOR

The main goal of the project was to create an XML tagged corpus of about 1000 French-English aligned normative texts from the Agence française de normalisation (*de jure* standard documents) in order to help the development of high scale linguistic applications such as translation tools, information tools for librarians, norm users and scientific researchers. Other outcomes of the project are:

- a tagging method that can be applied to other existing norms,
- a list of bilingual terms that may be used in technical dictionaries, as well as a set of phrases that can supply translation tools,
- search tools that use tag markers (search of the context of a term, the number of occurrences, exceptions in norms, etc.),
- prototypes including a multilingual semantic search engine and an automatic summarizer.

Project website: http ://stella.atilf.fr/dendien/recherche-afnor.htm

3.1.2. ALIZÉ

Project leader: ELISA consortium

ALIZÉ is a free software toolkit for speaker recognition developed by the ELISA consortium and designed to reach three objectives: being simple and easy to understand and to manipulate in order to be used by students for training and research purpose, being as efficient as possible according to the state-of-the-art, in terms of error rates but also in terms of computational resources needed, and facilitating the development of demonstrators or practical applications.

Project website: http ://www.lia.univ-avignon.fr/heberges/ALIZE/

3.1.3. OURAL

Project leader: Sinequa

The project developed three kinds of results: lexicons, corpora and tools which are disseminated with a GNU license. A wide range of lexicons have been created such as a French lexical database (160,000 forms, 44,000 lemmas, oral and written frequencies), a written text corpus of 37 million words, lexicons of anagrams, first names, homographs, etc. Different oral corpora have also been created (interviews concerning rental management of property with tenants, owners and interviewers and other contexts). Various tools have also been developed: for word processing, filtering documents, extracting concepts from documents and for automatic summarization.

3.1.4. WATSON

Project leader: Lingway

Watson has developed, adapted, integrated and/or generalized, language-based software tools for the logical structuring of web pages, the recognition of named entities, text marking, taggers, chunkers, extraction, categorization, co-reference resolution and summarization. The tools can be used separately or integrated into other solutions. They are designed to be robust and to offer optimal performance for tasks involving the processing of large volumes of data. A particular care has been taken to ensure their effective integration into web mining platforms. The modules are available for research and educational purposes by contacting the project leader.

3.1.5. NEOLOGOS

Project leader: TELISMA

The objective of the NEOLOGOS project is to provide two new speech databases (**Paidialogos** and **Idiologos**) that will be used in the framework of Automatic Speech Processing.

PAIDIALOGOS consists in the creation of a 1,000 speakers telephone database of children's voices, following the SpeechDat guidelines with some adaptations to the context of children speakers. The corpus contains various data types (numbers, sentences, command words, spelling...).

IDIOLOGOS is divided in 2 parts. The first part is a 1,000 speakers telephone database, each speaker having recorded 45 phonetically rich sentences in one call. The second part of the speech database is recorded by 200 speakers selected from the first part, each of them recording 450 sentences in 10 different calls containing more than 6,300 triphones. The objective of this second part was to create a speech database with reference speakers. Both, for the PAIDIALOGOS and IDIOLOGOS 1 databases, speakers have been recruited in the 12 French regions with a good distribution in terms of age and sex.

Within the **DICTIONNAIRE** cluster, four projects have been funded:

3.1.6. EURADIC

Project leader: CEA and CNRS

The goal of the **EurADiC** project was twofold. It first aimed at creating or completing various monolingual and bilingual dictionaries for common and specialized languages. Monolingual dictionaries for common languages have been developed for French, German, English, Spanish and Italian; the sizes of the resources balance between about 45,000 lemmas or parts of speech for the Italian dictionary to a little bit more than 155,000 entries for the English one. Bilingual dictionaries (for common language) have also been created for the following pairs for a minimum of 90,000 bilingual links: French-German, French-English, French-Arabic, French-Spanish and French-Italian. A terminological dictionary for French, English, German, Spanish, Greek and Arabic has been developed in the sport domain.

The second goal was to create a monolingual Arabic corpus and a bilingual French-Arabic corpus. The size of the first one is 105,000 words; this corpus is entirely voyelled and tagged. The second one includes 42 pairs of aligned texts at the sentence level. Both corpora come from *Le Monde Diplomatique*.

3.1.7. ATONANT

Project leader: EADS

The project aimed at developing a toolkit in order to facilitate the creation of semantic resources, namely ontologies. The toolkit is made of five independent tools ; the whole process is to crawl the web from specific URLs in order to constitute a corpus, to

normalize the information collected (TEI format), and to assist the creation of an ontology from the corpus. Each tool fits in a data processing sequence which allows processing from raw data (in this case Web pages) to a semantically enriched corpus. The toolkit has been developed and evaluated within the framework of a practical application, namely enrichment of resources in the medical domain.
Project website: http://atonant.insa-rouen.fr

3.1.8. Noms Propres

Project leader: University of Tours
The main goal of this project was to create multilingual resources for translation of proper names. These resources are not based on bilingual or multilingual dictionaries but on monolingual dictionaries sharing the same concepts and with interlingual links. To reach this goal, an ontology has been built which is structured in two parts: a monolingual part and a multilingual part. The core of the multilingual part is the concept of pivot which play the role of interlingual identifier. The database, which is available on line, exceeds 53,000 proper names (about 122,500 inflected forms) for French, with more than 400 relations of synonymy, more than 2,200 relations of accessibility, more than 44,000 relations of meronymy, and about 700 proper name translations in English, Italian, German, Spanish, Dutch, and Portuguese, and Serbian is presently considered (around 900 names).
Project website: http://tln.li.univ-tours.fr/tln_prolex/prolex.php

3.1.9. LEXITEC

Project leader: Softissimo
The **Lexitec** project led to the creation, by means of terminological extraction tools, of bilingual specialized dictionaries in domains where structured and easily exploitable lexical resources are lacking. The project has been implemented through the partnership of specialists in terminology, key actors in specialized domains and standard or technology prescribers. The results are dictionaries which are available in a format compatible for Automated Translation and standard exchange formats in the following domains: aerospace (6923 entries for FR⇒EN and 4271 entries for EN⇒FR), automotive (3382 for FR⇒SP and 2238 for SP⇒FR), business (5814 entries for FR⇒SP, 2016 entries for SP⇒FR), idiomatic expressions (4264 entries for FR⇒GE, 3766 for GE⇒FR, 1334 entries for FR⇒EN and 1064 entries for EN⇒FR) and mechanics (4554 entries for FR⇒EN, 3345 entries for EN⇒FR).

3.2. HLT Evaluation campaigns

Within the **EVALDA** cluster, eight evaluation campaigns have been organized:

3.2.1. ARCADE 2

The **ARCADE 2** project aims at exploring the techniques of multilingual text alignment through a fine evaluation of the existing techniques and the development of new alignment methods. The campaign consisted of two tracks devoted to the evaluation of alignment at sentence and word level respectively. The scenario of the sentence alignment task is defined as follows: given a set of parallel texts segmented into sentences, the participants had to return the alignment of sentences. There were two tracks in sentence alignment task: European language alignment and other languages alignment. For the sentence alignment task, multilingual reference corpora have been made available in 5 European languages and 6 languages of different writing systems. The word alignment exercise has involved a restricted task which is the identification of name entity translation in French-Arabic parallel texts.
For European languages, the JOC corpus has been used (about 5 million words equally spread in English, French, German, Italian and Spanish). The same subset for English, German, Italian and Spanish was aligned to their French counterpart at the sentence and paragraph level.
For non-European languages, a corpus of articles from the newspaper *Le Monde Diplomatique* has been used. It contains 150 Arabic texts aligned to French at the sentence level; 50 aligned text pairs with French as pivot language for Russian, Chinese, Japanese, Greek and Persian. A subset of French texts contains name entity tagged data for the word alignment task.

3.2.2. CESART

The **CESART** campaign concerns the evaluation of terminological resources acquisition systems. The project aimed at proposing and validating an evaluation protocol in order to compare different systems for terminology application such as terminological resource creation or semantic relation extraction. The project also aimed at creating high quality-controlled resources such as domain-specific corpora, automatic scoring tool, etc.
For the term extraction task, given a corpus of texts of a particular domain, participants were required to return a list of candidate terms ranked according to their relevance, with features concerning term frequency, variants and context information. An automatic process was first used to compare the output of the systems with a reference list generated from an existing thesaurus of the domain. All matched terms were considered as relevant. The non-matched terms were then submitted to experts for a relevance assessment.
For the semantic extraction task, besides the same corpus used in term extraction task, participating systems also received outputs of the same task. Participants are required to return a list of synonymy links between candidate terms. For this task, the overlap of the output of the systems against synonymy links extracted from an existing thesaurus was measured.
Two domain-specific corpora have been made available. The first one contains French documents collected from the web site of Health Canada (http://www.hc-sc.gc.ca) which lead to a 9 million word corpus. The second one is made of texts from SPIRALE, an education science journal, about half million word corpus. A third corpus, subset of French

texts extracted from the Written Questions and Answers of the Official Journal of the European Community (JOC), was used to mask the two test corpora and the outputs have not been evaluated.

3.2.3 CESTA

The **CESTA** project consisted in two machine translation (MT) evaluation campaigns involving English and Arabic source languages and French as the target language. The project first discussed different protocols and metrics traditionally used for MT evaluation such as the NIST/BLEU metrics, mWER and mPER. These metrics have been selected for the CESTA project and compared with other automatic evaluation metrics, based on the notions of grammatical score or semantic score such as X-Score, D-Score and WNM. The advantages and drawbacks of the various metrics have been studied within a meta-evaluation phase, prior to human evaluation.

For the first campaign, there were five participants for the EN→FR task and two participants for the Arabic→FR task. The English test corpus was composed of 15 documents belonging to the JOC corpus and the Arabic corpus was composed of 16 documents from the Unesco 32nd General Conference. In the 2nd campaign, 6 systems participated in the EN→FR task and one system in the Arabic→FR task. The corpus used for this campaign was a set of medical texts from the CESART project for the EN→FR task and a set of documents from the Unesco website together with articles from the *Al Hayat* newspaper for the Arabic→FR task. In both campaigns, the protocol consisted in embedding the test corpora in a larger corpus thematically homogeneous.

3.2.4. EASy

The aim of the **EASy** campaign was to design and test an evaluation methodology to compare syntactic analyzers on French and to produce a large validated language resource obtained by combining automatically the annotated corpora produced. The corpora consist of texts taken from various domains (literature, medicine, technical, general, etc.) and of different types (newspapers, questions, websites, speech transcriptions, etc.). The main results are a complete protocol of evaluation including corpora constitution, the manual corpora annotation, the evaluation of the participating parsers and the production of a large annotated resource (Treebank).

The evaluation measures adopted were the standard recall and precision methods which were calculated both on two tasks: evaluation of constituents and evaluation of dependencies. Nevertheless, since the corpora were heterogeneous, it was relevant to distinguish the results by the different types of corpus. Moreover, such separated results allow to know which parser is the best adapted to which situation.

In the same way it was relevant to evaluate the different subsets of relations separately. First, basic relations such as subject or direct object, secondly, the modifiers: noun modifiers, adjective modifiers... and lastly, relations that are less easy to calculate such as coordination, apposition or juxtaposition.

Five corpus providers participated in the campaign. They had to collect a large corpus of various genres (newspaper, medicine, literary texts…) and to annotate a part of it. This part (82,734 annotated words) has been "hidden" in the global corpus of 769,154 words. The evaluation process has been performed on the annotated part of the corpus but the 16 systems involved in the campaign had to annotate the whole corpus. One indirect result of the project is therefore the annotation of the global large corpus.

3.2.5. EQuerR

The project was designed to provide an evaluation framework for Question-Answering systems for the French language. **EQueR** included two tasks of automatic answering retrieval: the first one was a generic task carried out over a heterogeneous collection of texts, mainly newspaper articles, and the second one was a specialized task in the field of medicine carried out over a corpus of medical texts. To achieve these tasks, two corpora of respectively 500 and 200 questions have been worked out; they were made up according to the following 4 types of questions: "factual" (When did Kurt Cobain die?...), "definition" (What is IBM?...), "yes/no" (Did Jean-Paul II visit China?...) and "list" questions (Which are the 7 participating countries in the G7 group?...).

The evaluation phase of the systems took place at each participant site and lasted one week. For each question, the systems could return a short answer and/or a 250-character paragraph, provided that the short answers were not mandatory. The two types of answers were evaluated by two human judges and the results were provided on October 2004.

The project provided the following resources:

- French Generic Corpus (about 1.5 Gbytes), mainly newspaper articles,
- French Medical Corpus (about 140 Mbytes), texts extracted from several medical websites,
- an open domain set of 500 questions in French (generic field),
- a restricted set of 200 questions in French (medicine field),
- Generic and Medical sub-corpora from identifiers of documents extracted from the Pertimm search engine,
- a semi-automatic interface evaluation tool to help the human judges to evaluate the answers, the "EvalQA" tool (+ documentation),
- an automatic evaluation tool (under development): this tool will allow to evaluate answers and score the evaluation results without any human intervention.

3.2.6. ESTER

The **ESTER** campaign aims at radio-broadcast news transcription systems performance evaluation. The transcriptions are enriched by a set of side information, such as the segmentation into speech turns, the marking of named entities, etc The evaluation of the performance on this side information in addition to the evaluation of the orthographic transcription itself allows to establish a

reference for the present state-of-the-art of each component of an indexing system, while providing a global information on the performances of the complete systems. 13 laboratories participated in the evaluation, including 3 companies. Results were reported on written transcriptions in real-time or non real-time, on segmentation (speech/sound, speaker recognition and speaker diarization), and on named entity recognition, either directly from speech or from the written transcription). Very large language resources have been produced in the framework of the ESTER evaluation campaign: 100 hours of manually transcribed speech, corresponding to one million words and 350 speakers, and 1600 hours of untranscribed speech. An evaluation package will be distributed, including the development and test data, the scoring software and the results. Part of the data has also been used in the EASy syntactic parser evaluation campaign. A final workshop took place in march 2005, in order to report results. Another workshop was organized for the linguists in may 2005, where the data available and the results were presented, stressing the bottlenecks which would require further basic research investigations.

3.2.7. EVASY

The **EVASY** project is an evaluation campaign of text-to-speech synthesizers (TTS) in French. The project is a follow-up of the AUF TTS evaluation action (1996-99), which was one of the very few evaluation experiences ever conducted on that topic for the French language. The evaluation campaign is divided in four parts: evaluation of grapheme-to-phoneme (GP) conversion, prosody, intelligibility test and global quality of the synthesized speech.

Four systems participated in the GP conversion task which consisted in phonetizing a list of proper names within 3 hours, four systems in the prosody task and five systems in the global TTS quality task. The GP evaluation took place in December 2004 on the NIST metrics basis. For this task which concentrated on proper names as it has been shown that the majority of GP errors occurred from them, a list of 8,230 names, extracted from *Le Monde* has been created, and hand-transcribed in the phonetic alphabet SAMPA for the French language. Additionally, the list has been enriched with linguistic origin indications concerning the surnames.

The prosody evaluation took place in October 2005; the outputs of the different systems have been pasted on the same diphones base in order to get the same voice for all the sentences. A corpus of seven sentences from the BREF corpus has been created and the results have been evaluated by 19 assessors.

The intelligibility test involved 6 systems and was based on the SUS paradigm (Semantically Unpredictable Sentences), which allows for an objective assessment of intelligibility at word-level. A list of 288 SUS has been built, divided into blocks of 4 different syntactic structures.

Finally, the global quality evaluation used the Absolute Category Rating (ACR). In addition to the MOS (Mean Opinion Score) metric, six other criteria have been adapted to the French language from the Verbmobil project. The corpus used for the evaluation was EUROM 1, developed within the Multext and Esprit SAM projects.

3.2.8. MEDIA

The aim of the **MEDIA** project is to define and experiment an evaluation methodology for dialog systems. The evaluation paradigm is based on the definition and use of a battery of tests extracted from real corpora, with common semantic representations and metrics. This should allow for diagnosing the capacity of the dialog systems for context-independent and context-dependent understanding. The evaluation campaign has been conducted in order to validate in a first step the paradigm and the representations on a common task related to information query. Five systems from academic labs and based on very different models, participated to the evaluation. A corpus of dialogs has been produced using a Wizard-of-Oz approach and has been semantically annotated. This task has been processed with a specific annotation tool in agreement with a semantic dictionary defined for the project.

The same annotated training corpus was given to each participant in order to adapt its own model to the task and the domain, as well as the semantic dictionary, the annotation manual and the 3,786 words lexicon from the Media corpus.

After a dry-run in April 2005 on a 1,000 utterances set, the real campaign was performed in June 2005 on a test set of 3003 unknown utterances randomly extracted from the corpus of dialogs.

The results provided by the systems have been scored by a specific tool able to align and compare the semantic representations of the queries.

The resources obtained from the project are mainly a semantic dictionary including 83 basic attributes and 19 specifiers, and a dialog annotated corpus.

3.3. Standardization

Within the **NORMALANGUE** cluster, two projects have been funded, concerning spoken and written language.

First, the goal of the **RNIL** project was to help French companies to participate in the elaboration of *de jure* standards at the international level, the ISO TC37 SC4 which is responsible for language description. A national committee has been established under the auspice of AFNOR (the French official body for standardization) both for collecting positions and comments from companies and academic labs and diffusing information coming from the various international groups. Basically, the working fields opened by the project concern the *Morpho-Syntactic Annotation Framework*, the *Lexical Model Framework* and the *Data Category Registry*. These works are still in progress but significant results have

already been reached. A big issue of the project is to widespread standards to companies in order to help them in the changing normative environment.
The goal of the other project, **Technovox**, was very similar but in the speech domain. Much work has been done to enrich the VoiceXML and the UNL standards. Due to the specific nature of speech community and the very rapid technological progress in the field, the standardization process takes place in various informal groups and forums. Typically, the project gave the possibility to participate in the VoiceXML Forum of the W3C, the AURORA committee from ETSI, the IETF, the SALT Forum or the UNL Foundation. As for the RNIL project, the other important goal is to widely disseminate the information coming from these bodies.

3.4. Information survey

The last goal of the Techno-langue program was to create a portal dedicated to HLT. This portal is accessible at the following address: www.technolangue.net. It has been established in relation with the different associations operating in that field, such as APIL (the French Professional Association in HLT), AFCP (the Academic Society for Spoken Language) and ATALA (the Academic Society for Natural Language Processing). Its goal is to make available different kinds of resources such as news of the field, market studies, success stories, various directories of laboratories, companies, institutes, glossaries, etc.

4. Towards a European infrastructure

Taking into consideration the fact that national programs in the field of HLT have been recently launched in several European Member-States and the opportunity given by the European Commission Framework Program to coordinate national programs, it has been proposed to launch an initiative in agreement with the European Research Area concept.
Language is a specific issue for Europe. It is mandatory to address this issue at the same time to preserve the culture, and therefore languages of each Member States, and to facilitate the communication among the European citizens speaking different languages. The charge is especially tedious for the European Commission and for the European Parliament, which devote a large part of their budget to take care of the 20 official European languages, corresponding to 380 language pairs. Providing Language Technologies would enlighten this charge, but it appears that the task is too heavy for the European Commission alone, and would benefit from sharing the effort between the EC and the Member-States. The EC would primarily take into account what is independent of a specific language, coordination, management, standards, core technology development and evaluation, communication and awareness, while the Members States would primarily take into account what is specific to their language(s): Language Resources and language technology adaptation.
This initiative, called *Lang-Net*, addresses the following strategic objectives:

- to coordinate activities in the field of LT among European countries,
- to ensure a proper coverage of Language Resources and Technologies for the various languages spoken in those countries,
- to prepare a large program on LT as a synergy of efforts between the European Commission and the European Member States, in the perspective of the 7th Framework program, using the proper instruments (ERA-Nets, Article 169 or JTI),
- to guarantee the availability of the needed LT in order to facilitate communication among the European citizens within a multilingual European Union.

The Lang-Net proposal presently gathers 15 partners from 11 national or regional governments: Basque Region, Czech republic, Denmark, France, Germany, Italy, Norway, Spain, Sweden, The Netherlands / Belgium Flanders, and Province of Trento.

5. Conclusion

As it has been showed, the *Techno-langue* program allowed for the creation of different kinds of high-quality language resources and basic toolkits for spoken and written language, both in the specific action line and in the evaluation action line. Even before the formal evaluation of the program, we may state that it has reached this first aspect of the original goals. The second important aspect which is still in discussion with some project leaders, is the legal and financial conditions of the distribution of these resources. It has not been possible, neither relevant, to constrain all the projects to respect the same distribution model. The relevant approach is a very pragmatic one, based on a case-by-case discussion. The only point which is not negotiable is that the resources must be made available to the HLT academic and industrial community.
Concerning the *Evalda* platform, all the 8 campaigns reached their objectives of evaluating technologies and creating specific resources for evaluation. It is planned to distribute "evaluation packages" for all of them. Another result of big interest has been to demonstrate both to academics and to policy makers the importance of technology evaluations in the innovative process.
Our task is now to present the results to the scientific community, to industrials and to administrations, to settle a permanent infrastructure which can feed applications with validated technologies and to promote a similar infrastructure at the European level within FP7, as a large, coordinated and shared HLT program.

6. References

See the Techno-langue web site at: www.technolangue.net to access the whole bibliography concerning the various projects.