Jornada de Seguimiento de Proyectos, 2005
Programa Nacional de Tecnologías Informáticas

TEFATE: Finite-state technologies applied to specialized translation*
TIC2003-08681-C02

Mikel L. Forcada †
Dept. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant

Francisco Casacuberta Nolla ‡
Dept. de Sistemes Informàtics i Computació
Universitat Politècnica de València

Abstract

In the current European scenario, characterized by the coexistence of communities writing and speaking a great variety of languages, machine translation has become a technology of capital importance. In areas of Spain and of other countries, coofficiality of several languages implies producing several versions of public information. Machine translation between three languages of the Iberian Peninsula and from them into English will allow for a better integration of Iberian linguistic communities among them and inside Europe. This project aims at applying the know-how developed by the participating teams in previous projects dealing with general machine translation between two languages (Spanish and Catalan) and specific machine translation between several languages (Spanish and Italian into English) to develop *finite-state technologies* for translation and to apply them to build tools to efficiently harness these methods for new language pairs (Spanish–Portuguese, Catalan–English) or subjects (weather reports, tourist information, medical information, phone banking, etc.). We are applying the finite-state technologies developed to (a) the creation of *machine translation systems* for texts and *interactive translation assistance systems* in specialized subjects based upon finite-state technologies from existing machine translation dictionaries and information obtained from *bitexts* (parallel bilingual texts) harvested from the web; (b) the construction of easily-maintained, fast, web-based, finite-state *translation memories* from bitexts; (c) the production of *voice-to-voice translation systems* for specific subjects, trained on specialized bitexts; and (d) the creation of databases of aligned, morphosyntactically annotated *bitexts* for future research on translation technologies. We have built *harvesting robots* to collect bitexts from the web, statistical and geometrical *aligners* enriched with preexisting linguistic information and text and bitext *classifiers* to delimit specific translation subjects.

*Spanish title: "Tecnologías de estados finitos aplicadas a la traducción especializada"

†Email: mlf@ua.es
‡Email:fcn@iti.upv.es

Keywords: finite-state transducers, text-to-text machine translation, speech-to-speech machine translation, statistical machine translation, classifiers, parallel corpora.

1 Project objectives

This project has, as a basic objective, the development of a range of *finite-state technologies* for translation and, as applied objectives,

1. creating *machine translation systems* and *interactive translation assistants* for texts in specialized subjects, based on the translation dictionaries already built by the groups and on information obtained from *bitexts* (bilingual parallel texts) harvested from the Internet;
2. building easily-maintained fast-response *translation memories* which may be used through the internet, also based on finite states, and are fed with bitexts harvested from the Internet;
3. producing *speech-to-speech* machine translation systems for specialized applications, trained on specific bitexts; and
4. building databases of aligned, morphosyntactically annotated bitexts as resources for future research.

The project will tackle the new language pairs Spanish–Portuguese, Spanish-Basque and Catalan–English, and, among other specialized subjects, tourist information, meteorological reports, patient–doctor dialogs, and phone banking.

2 Degree of performance

2.1 Activity and achievements

This section describes the activity for each module (M) and task (T) of the project. The name of the university (UA, UPV) leading each task appears in parenthesis, together with scheduling details.

M1 Task definition and coordination

T1 Project coordination (UA: 12/2003–11/2006): three coordination meetings have been held: one informal kick-off meeting and two workshops (November 2004 and May 2005); a webpage has been set up (`http://transducens.dlsi.ua.es/TEFATE/` to coordinate the project). We are trying to improve the coordination between groups and the content of the project webpage.

T2 Definition of tasks for text-to-text and speech-to-speech translation (UA, UPV: 12/2003–2/2004).

- For the text-to-text tasks, we have chosen the following language pairs: Spanish–Catalan (corpora: *Diari Oficial de la Generalitat Valenciana* y *El Periódico de Catalunya*) and Spanish–English (corpora: Xerox printer manuals, session minutes of the European Union).

- For the speech-to-speech tasks we chose the following language pairs: Catalan–English and Portuguese–English; in both cases, the task chosen was the tourist task (sentences said by a tourist at a hotel reception desk). In another project related to this one, a similar corpus was defined for a Basque–Spanish translation task.

M2 Acquisition of multilingual texts from the Internet.

T1 Acquisition of multilingual texts from the Internet (UA, UPV: 12/2003–11/2004).

- Using ad-hoc robots, the Valencia group has harvested the following corpora: *Diari Oficial de la Generalitat Valenciana* (DOGV, 19.5 million words in Spanish, 19 million in Catalan, vocabulary sizes 189,000 and 197,000 respectively) and *El Periódico de Catalunya* (14.4 million words in Spanish and 15 million words in Catalan, vocabulary size: 126000 words each).
- Development of tools for the automatic harvesting of corpora from the Internet. A general-purpose robot has been developed and is currently being tested, which harvests bitexts in five languages (Spanish, Catalan, Portuguese, Galician and English) starting from a list of seed URIs; the robot is able to identify the languages of pages and relating them to their translations.
- Elaboration of a new multi-reference bitext for the evaluation of the dictionaries and morphopsyntactical taggers used.

T2 Bitext validation (UA: 3/2004–2/2005). The general-purpose harvesting robot being tested evaluates a set of different features of harvested bitext candidates, and delivers only those bitexts above some predefined thresholds.

T3 Data filtering (UA, UPV: 6/2004–5/2005).

- Filtering of harvested data (DOGV, *El Periódico*).
- Automatic tagging and manual testing of corpora for use with a semantic disambiguation module.
- Development of an encoding scheme for the semantical annotation of translation units.
- The threshold parameters used in the general-purpose harvesting robot will be defined through human validation of samples of harvested bitexts using a visual tool.

T4 Integration of linguistic data and structural information in statistical aligners (UA: 9/2004–8/2005). The project has developed a class of aligners based on linguistic-independent heuristics (LIHLA, [3, 4, 5]) with interesting results; however, advancement in aligning methods which use finite-state transducers has been insufficient.

M3 Document classification.

T1 Training of plain text document classifiers (UPV: 6/2004–5/2005).

- Development of finite-state-based classifiers, as well as classifiers based in an error-correcting adaptation of the Viterbi method.
- Development of classifiers based on statistical models, such as multinomial mixtures, as well as their specific estimation methods (expectation maximization).

- Application to text classification (in the Tourist task, the classification error has been improved by 2%).

T2 Structural classification of hypertexts (UA: 12/2004–11/2005). This task has not been initiated (waiting for validated data obtained using the general-purpose harvesting strategies). A method will be implemented, based on probabilistic tree grammars [1] to classify XHTML bitexts.

M4 Building of finite-state and statistical translators.

T1 Improvement of dictionary-based finite-state translators (UPV: 12/2003–5/2006)

- Application of statistical disambiguation using naïve Bayes multinomial methods.
- Improvement of the translation speed of the TAVAL/SISHITRA machine translation system (up to 500 words per second) through a complete rewriting.
- Debugging, extension and improvement of the TAVAL/SISHITRA dictionary through the introduction of complex, inflected translation units, frequent errors, specific terminology and prepositional regime.
- Complete rewriting of the Alacant shallow-transfer machine translation engine (used in the interNOSTRUM and Traductor Universia systems) for release under a GPL license (Corbí-Bellot et al. 2005, `http://sourceforge.net/projects/apertium/`), in collaboration with another publicly-funded project.

T2 Inference of specialized finite-state transducers (UPV: 5/2004–11/2005).

- Incorporation of statistical models into the GIATI technique [2].
- Improvement of smoothing techniques.
- Application of transducer composition.
- Development of multiple translation techniques.

T3 Training of general statistical translators (UPV: 12/2004–5/2006).

- Implementation of statistical Spanish–Catalan translators.
- Using TAVAL/SISHITRA dictionaries to improve the statistical translators.
- Development of automatic techniques for the selection of word parts.
- Developing statistical-model combinations using translation error minimization as a criterion.

T4 Translator smoothing (UPV: 12/2005–11/2006): not started yet.

M5 Subsentential finite-state translation memories.

T1 Incremental building finite-state transducers from translation memory files (UA: 3/2004–2/2005).

- Development of a very fast compiler that takes a complete translation memory and generates a finite-state transducer which is read by a module capable of treating tens of thousands of words per second (Ortiz-Rojas et al. 2005). The compiler is so fast that incrementality may not be considered an advantage, although it has to be tested on translation memories harvested from the Internet in Module 2.

- Design of a new algorithm (Recalign) for the recursive bilingual segmentation of bitexts using a greedy strategy and IBM model 1 (UPV, even if task assigned to UA).

T2 Efficient implementation of finite-state translation memories (UA: 6/2004–5/2005, this task has been merged with M5T1, q.v.).

T3 Inference of left-to-right, longest-match translation memories from partially aligned texts (UA: 6/2005–5/2006, just started).

M6 Systems for assisted translation

T1 Finite-state based assisted translation (UPV: 6/2005–11/2006, just started)

T2 Statistical machine translation based assisted translation (UPV: 6/2005–11/2006, just started)

M7 Creation of syntactically-analized bitext corpora.

T1 Building of probabilistic analyzers with partial relaxation of independence constraints (UA: 12/2004–11/2005). Task in progress. An implementation has been recently described [8].

T2 Design of systems to assist in the syntactical annotation of bitexts (UA: 12/2005–11/2006, not started yet).

M8 Building of translation servers

T1 Building of speech-to-speech translation prototypes for specific tasks (UPV: 12/2004–5/2006).

- Building the Catalan–English, Portuguese–English and Spanish–Basque translators for the Tourist task.

T2 Implementation of text-to-text translators as a server (UA: 6/2006–11/2006, not started yet).

T3 Implementation of the user interface (UA: 6/2006–11/2006, not started yet).

2.2 Difficulties and proposed solutions

2.2.1 Subproject TIC2003-08681-C02-01: TEFBARNet, UA

There have been some problems related to the internal restructuring of the research group, which has delayed some of the tasks. Adjustment to the programmed schedule will require a redistribution of tasks as well as a special effort of the active researchers.

2.2.2 Subproject TIC2003-08681-C02-02: ITEFTE, UPV

The main problems encountered during the development of the project have been mainly related to the acquisition, filtering, alignment and categorization of corpora, which has ended up having a higher temporal cost than what was initially planned. The multiple morphosyntactical annotation of translation units inside the dictionaries has also proven to be a slow process.

3 Result indicators

3.1 Trainees

3.1.1 Subproject TIC2003-08681-C02-01: TEFBARNet, UA

The project has a single trainee, Felipe Sánchez Martínez, who has a four-year scholarship to work on a thesis related to the project [6, 7].

3.1.2 Subproject TIC2003-08681-C02-02: ITEFTE, UPV

- Jesús Andrés Ferrer, computer engineer, collaborating scholarship holder (assigned to the project since January 2005). Work area: development of new translation models.
- Laura María Eliodoro Furió, philologist, collaborating scholarship holder (until May 2005). Work area: development, annotation and correction of dictionaries.
- Adrián Giménez Pastor, computer engineer, collaborating scholarship holder. Work area: acquisition of corpora.
- Maria Teresa González Cavero, computer engineer, scholarship holder assigned to the project. Work area: development of translation assistance tools.
- Antonio Luis Lagarda Arroyo, computer engineer, scholarship holder assigned to the project. Work area: development of translation assistance tools.

3.2 Employees

3.2.1 Subproject TIC2003-08681-C02-01: TEFBARNet, UA

The following technical stuff has been hired as *técnico superior*:

- Sergio Ortiz Rojas (February 2004–February 2005), computer engineer. Work area: transducer compilers, translation memories.
- Enrique Sánchez Villamil (February 2004–), computer engineer. Work area: bitext harvesting robots.
- Susana Santos Antón (February 2005–), computer engineer. Work area: validation tools for harvested bitexts.

3.2.2 Subproject TIC2003-08681-C02-02: ITEFTE, UPV

The following technical stuff has been hired as *técnico superior*:

- Laura María Eliodoro Furió, philologist (from May 2005). Work area: development, annotation, and correction of dictionaries.
- José García Hernández, computer engineer. Work area: corpora acquisition.
- Luis Rodríguez Ruiz, computer engineer. Work area: development of tools for assisted translation.

3.3 Technology transfer

3.3.1 Subproject TIC2003-08681-C02-01: TEFBARNet, UA

Many of the basic techniques developed during this project, especially those related to finite-state transducers, have been transferred to Apertium, an open-source shallow-transfer machine translation engine scheduled to be released at the end of July 2005 and developed as part of another project. Apertium will be used by Imaxin, a software company, to market machine translation services for the Spanish–Galician pair.

3.3.2 Subproject TIC2003-08681-C02-02: ITEFTE, UPV

Although in principle the collaboration of external companies in the subproject was not programmed, some companies —with which contacts had already be maintained during past projects or collaborations— have shown their interest in the results of some of the tasks developed. It is the case of ATOS Origin, who specialize in the development of tools for assisted translation; Celer Soluciones, a technical translation company; Xerox Research Center Europe (XRCE), in view of their needs to incorporate tools to translate technical manuals; and ASP (Adur Software Productions), who are interested in improving their translation-memory-based tools.

It is also the case that some of the results of this project have been useful in other technological development projects with enterprises, such as Rumbo Sistemas and ODEC.

3.4 Participation in international projects

3.4.1 Subproject TIC2003-08681-C02-02: ITEFTE, UPV

The European project TT2, in the IST program, in which the UPV group also participates, has benefited from the results obtained in the project, which have been applied in the last period of the TT2 project. The success of the applied techniques and the results obtained have prompted the groups to apply for two more projects in the fifth call of the EU FP6. Both applications have a strong relation to the subject of the present project.

3.5 Publications

3.5.1 Journal papers

- I. García-Varea and F. Casacuberta. Maximum entropy modeling: A suitable framework to learn context-dependent lexicon models for statistical machine translation. Machine Learning, 59:1–24, 2005.
- Verdú-Mas, J.L., Carrasco, R.C., Calera-Rubio, J. (2005). Parsing with probabilistic strictly locally testable tree languages. IEEE Trans. on Pattern analysis and Machine Intelligence, 27(7):1040–1050.
- E. Vidal, F. Thollard, F. Casacuberta C. de la Higuera, and R. Carrasco. (2005) Probabilistic finite-state machines - part I. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7):1013–1025.

- E. Vidal, F. Thollard, F. Casacuberta C. de la Higuera, and R. Carrasco. (2005) Probabilistic finite-state machines - part II. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(7):1025–1039.
- F. Casacuberta, E. Vidal, and D. Picó. (2005) Inference of finite-state transducers from regular languages. Pattern Recognition, 38:1431–1443.

3.5.2 Papers at conference proceedings

- J. Andrés, J. Navarro, A. Juan, and F. Casacuberta (2005). Word translation disambiguation using multinomial classifiers. In Iberian Conference on Pattern Recognition and Image Analysis, volume 3523 of Lecture Notes in Computer Science, pages 622-629. Springer-Verlag, Estoril (Portugal), June 2005.
- Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. (2005a, aceptado) LIHLA: A lexical aligner based on language-independent heuristics. En V Encontro Nacional de Inteligência Artificial (ENIA 2005). São Leopoldo-RS, Brasil, julio 2005.
- Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. (2005b, aceptado). Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Granada, setiembre 2005.
- Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. (2005c, aceptado). LIHLA: Shared task system description. Workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond (ACL05). Michigan, EE.UU., junio 2005.
- J. Civera, J.M. Vilar, E. Cubel, A.L. Lagarda, S. Barrachina, F. Casacuberta, E. Vidal, D. Picó and J. González (2004). A syntactic pattern recognition approach to computer assisted translation. In A. Fred, T. Caelli, A. Campilho, R. P.W. Duin and D. de Ridder, editors, Advances in Statistical, Structural and Syntactical Pattern Recognition, Lecture Notes in Computer Science. Vol 3138, pp. 207-215 Springer-Verlag.
- Corbí-Bellot, A.M. Forcada, M.L., Ortiz-Rojas, S. Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., Sarasola, K. (2005). "An open-source shallow-transfer machine translation engine for the romance languages of Spain", Proceedings of the 10th Annual Conference of the European Associatation for Machine Translation, p. 79-86, 30-31 mayo 2005, Budapest, Hungría.
- I. García-Varea, A. Sanchis and F. Casacuberta (2004). A decoding algorithm for speech input statistical translation. In Text, Speech and Dialogue: Proceedings of the 7th International Conference (TSD 2004), volume 3206 of Lecture Notes in Computer Science, pages 305-314. Springer-Verlag, Brno, Czech Republic, September 2004.
- I. García-Varea, D. Ortiz, F. Nevado, P.A. Gómez, and F. Casacuberta. (2005) Automatic segmentation of bilingual corpora: A comparison of different techniques. In Iberian Conference on Pattern Recognition and Image Analysis, volume 3523 of Lecture Notes in Computer Science, pages 614-621. Springer-Verlag, Estoril (Portugal), June 2005.
- J. González, D. Ortiz, J. Tomás and F. Casacuberta (2004). A comparison of different statistical machine translation approaches for spanish-to-basque translation. In Actas de las Terceras Jornadas de Tecnología del Habla, Valencia, Spain, November 2004.

- C. D. Martínez-Hinarejos, L. Rodríguez, J. A. Sánchez, A. Sanchis and E. Vidal (2004). Comparison of several speaker adaptation alternatives for a spanish speech task. In Actas de las III Jornadas de Tecnologías del Habla, Valencia, Spain, November 2004.
- J. R. Navarro, J. González, D. Picó, F. Casacuberta, J. M. de Val, F. Fabregat, F. Pla and J. Tomás (2004). SisHiTra: A hybrid machine translation system from Spanish to Catalan. In J. L. Vicedo, P. Martínez-Barco, R. Muñoz and M. Saiz, editors, Advances in Natural Language Processing, Lecture Notes in Artificial Intelligence, pages 349-359. Springer-Verlag, Alicante, Spain.
- F. Nevado and F. Casacuberta (2004). Bilingual corpora segmentation using bilingual recursive alignments. In Actas de las III Jornadas en Tecnologías del Habla, 3JTH, Valencia, November 2004.
- F. Nevado, F. Casacuberta and J. Landa (2004). Translation memories enrichment by statistical bilingual segmentation. In Proceeedings of the IV International Conference on Language Resources and Evaluation - LREC2004, volume 1, pages 335-338, Lisbon, May 2004.
- Ortiz-Rojas, S., Forcada, M.L., Ramírez-Sánchez, G. (2005). Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Granada, setiembre 2005.
- D. Picó, J. Tomás and F. Casacuberta (2004). GIATI: A general methodology for finite-state translation using alignments. In Statistical, Structural and Syntactical Pattern Recognition. Proceedings of the Joint IAPR Intternational Workshops SSPR2004 and SPR2004, volume 3138 of Lecture Notes in Computer Science, pages 216-223. Springer-Verlag, Lisboa, Portugal, August 2004.
- Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. (2004a). Exploring the use of target-language information to train the part-of-speech tagger of machine translation systems. Lecture Notes in Computer Science Lecture Notes in Artificial Intelligence 3230 (Advances in Natural Language Processing, Proceedings of EsTAL - España for Natural Language Processing), p. 137-148, 20-22/10/2004, Alicante, España.
- Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. (2004b). "Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system". Proceedings of TMI 2004, The Tenth Conference on Theoretical and Methodological Issues in Machine Translation, p. 135-144, 4-6/10/2004, 2004, Baltimore, MD, EE.UU.
- Sánchez-Villamil, E.; Forcada, Mikel L.; Carrasco, Rafael C. "Unsupervised Training of a Finite-State Sliding-Window Part-of-Speech Tagger" (2004). Lecture Notes in Computer Science - Lecture Notes in Artificial Intelligence 3230 (Advances in Natural Language Processing, Proceedings of EsTAL - España for Natural Language Processing), p. 454–463, 20-22/10/2004, Alicante, España.
- J. Tomás and F. Casacuberta (2004). Statistical machine translation decoding using target word reordering. In Statistical, Structural and Syntactical Pattern Recognition. Proceedings of the Joint IAPR Intternational Workshops SSPR2004 and SPR2004, vol-

ume 3138 of Lecture Notes in Computer Science, pages 734-743. Springer-Verlag, Lisboa, Portugal, August 2004.

- J. Tomás, J. Lloret and F. Casacuberta (2004). A Spanish-Catalan translator using statistical methods. In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI04), pages 1099-1100, Valencia, Spain, August 2004. IOS Press.
- J. Tomás, J. Lloret and F. Casacuberta (2004). Spanish-catalan translator using statistical methods. In Actas de las Terceras Jornadas de Tecnología del Habla, Valencia, Spain, November 2004.
- J. Tomás, J. Lloret, and F. Casacuberta (2005). Phrase-based alignment models for statistical machine translation. In Iberian Conference on Pattern Recognition and Image Analysis, volume 3523 of Lecture Notes in Computer Science, pages 605-613. Springer-Verlag, Estoril (Portugal), June 2005.
- E. Vidal and F. Casacuberta (2004). Learning finite-state models for machine translation. In Grammatical Inference: Algorithms and Applications. Proceedings of the 7th International Coloquium ICGI 2004, volume 3264 of Lecture Notes in Artificial Intelligence, pages 16-27. Springer, Athens, Greece, October 2004. Invited conference.

3.6 Activities planned until the end of the project

- Text classification:
  - Classifying bilingual texts using mixture models: combination of IBM models and n-grams. (UPV).
  - Developing hypertext classifiers based on their structure, using probabilistic grammars (UA).
- Development of aligners and translators:
  - Transforming statistical models into finite state models (UPV)
  - Combining GIATI [2] with other statistical models (UPV)
  - Using linguistic information in the automatic construction of statistical and finite-state translators (UPV)
  - Deploying language-independent heuristic aligners (LIHLA, [3, 4, 5]) to obtain linguistic data for machine translation systems (UA).
  - Implementing geometrical aligners based on edit costs which incorporate linguistic information through finite–state transducers (UA).
- Developing search algorithms for statistical translators:
  - For mixtures of IBM models (UPV).
  - Multiple statistical models (UPV).
- Development and debugging of dictionaries and linguistic data:
  - Improving of the TAVAL dictionary by introducing semantic information (UPV).

  - Automatically inducing the topology (tag groupings) and parameters of part-of-speech taggers using information from the target language (UA)
  - Exploring the use of simple, target-language driven statistical choice models to treat polysemic words in rule-based machine translation systems (UA).

- Implementation of machine translation systems:
  - Improving the code of SisHiTra-TAVAL. Augmenting SisHiTra-TAVAL with the capability to translate web pages (UPV [UA]) and LaTex files.
  - Implementing the Catalan–Spanish version of SisHiTra-TAVAL (UPV).
  - Implementing systems for multiple translation (UPV).
  - Extension of the Apertium (interNOSTRUM–Traductor Universia) architecture to deal with deeper syntactic transfer and multiple lexical equivalents for polysemic words (UA).

- Assisted translation/transcription systems:
  - Building interactive assisted translation systems: interface, translation engines and incorporation of the SisHiTra–TAVAL system(UPV)
  - Exploring the development of systems for assisted trancription (UPV).

- Harvesting of bitexts from the Internet
  - Use of a bitext evaluation tool by humans to tune the parameters (threshold) of the automatic bitext harvesting robot (UA).
  - Exploring conservative techniques to obtain "safe" subsentential units from fully-automated Internet-harvested bitexts (UA).
  - Creation of a finite-state based prototype that will manage the harvested translation units to act as a translation assistance system (UA).

- Creating syntactically-analysed bitext corpora:
  - Developing tools for the simultaneous syntactic annotation of bitexts using online induction of probabilistic grammars (UA).

3.7 Collaboration with other groups

3.7.1 Subproject TIC2003-08681-C02-01: TEFBARNet, UA

As part of the project activity, a collaboration has started with the Núcleo Interinstitucional de Lingüística Computacional (NILC) of the Universidade de São Paulo (USP) en São Carlos (SP), led by professsor Maria das Graças Volpe Nunes. As part of this collaboration,

- Helena de Medeiros Caseli, a student from that group, has visited the Universitat d'Alacant for one year, funded by a Brazilian agency. Helena has been developing a class of bitext aligners, and has produced three publications [3, 4, 5].
- The principal investigator has paid a short visit to the Brazilian group related to the work of this student.

Also, a collaboration has been initiated with Prof. Jan Daciuk, of the Department of Knowledge Engineering of the University of Gdańsk, in Poland. The subject of this collaboration is the incremental construction of minimal tree automata (Prof. Daciuk visited the UA in October 2004).

3.7.2 Subproject TIC2003-08681-C02-02: ITEFTE, UPV

The project has produced an important collaboration with the Portuguese group *Laboratório de Sistemas de Língua Falada* of the Instituto Superior Técnico in Lisbon. This collaboration has been organized around an Integrated Action between both groups, which has allowed mutual visits and an intense collaboration in research tasks.

References

[1] R.C. Carrasco, J.R. Rico-Juan, "A similarity between probabilistic tree languages: application to XML document families", Pattern Recognition 36:9 (2002) 2197-2199.

[2] F. Casacuberta "Inference of finite-state transducers by using regular grammars and morphisms", Lecture Notes in Computer Science, vol. 1891, p. 1–14.

[3] Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. LIHLA: A lexical aligner based on language-independent heuristics. En V Encontro Nacional de Inteligência Artificial (ENIA 2005). São Leopoldo-RS, Brasil, julio 2005.

[4] Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. XXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Granada, setiembre 2005.

[5] Caseli, H.M.; Nunes, M.G.V.; Forcada, M.L. . LIHLA: Shared task system description. Workshop on Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond (ACL05). Michigan, EE.UU., junio 2005.

[6] Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. Exploring the use of target-language information to train the part-of-speech tagger of machine translation systems. Lecture Notes in Computer Science Lecture Notes in Artificial Intelligence 3230 (Advances in Natural Language Processing, Proceedings of EsTAL - España for Natural Language Processing), p. 137-148, 20-22/10/2004, Alicante, España (2004).

[7] Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. "Cooperative unsupervised training of the part-of-speech taggers in a bidirectional machine translation system". Proceedings of TMI 2004, The Tenth Conference on Theoretical and Methodological Issues in Machine Translation, p. 135-144, 4-6/10/2004, 2004, Baltimore, MD, EE.UU.

[8] J.L. Verdú-Mas, R.C. Carrasco, J. Calera-Rubio. Parsing with probabilistic strictly locally testable tree languages. IEEE Trans. on Pattern analysis and Machine Intelligence, 27(7):1040–1050 (2005).