A 45-hour Computers in Translation course

Mikel L. Forcada
Departament de Llenguatges i Sistemes Informtics
Universitat d'Alacant, E-03071 Alacant, Spain.
mlf@ua.es

Abstract

This paper describes how a 45-hour Computers in Translation course is actually taught to 3rd-year translation students at the University of Alacant; the course described started in year 1995–1996 and has undergone substantial redesign until its present form. It is hoped that this description may be of use to instructors who are forced to teach a similar subject in such as small slot of time and need some design guidelines.

1 The subject

The subject Computers in Translation (officially *Informática Aplicada a la Traducción*) is a mandatory subject in all official 4-year or 5-year translation degrees in Spain, and can be taught as part of any of the last two years, with a minimum of 4.5 credits (in Spain, a credit is equivalent to 10 h of classroom or laboratory time but can be reduced in some circumstances). The official Ministry of Education descriptor for the subject is very short:

> Access to the necessary tools for translation work. Machine translation and computer-assisted translation. System integration.

This leaves a lot of freedom to design the actual syllabus. Some universities extend these 4.5 credits and split the subject in two or more courses; other universities add mandatory ”computer literacy” courses during the first two years. The University of Alacant simply satisfies the minimum requirements: a single 45-h Computers in Translation course is programmed as part of the 3rd year; this course is expected to provide future translators with enough knowledge and skills about the application of computers to translation.

2 Students, groups and sessions

In Alacant, we have about 150 students each year from German (30), English (60) and French (60) translation studies, who meet in either of two 75-student classroom groups and in one of six 25-student laboratory groups, regardless of the language. The course is organized in 30 1.5-hour sessions (15 weeks, 2 sessions a week): 19 sessions are held in the classroom and 11 in a computer laboratory. In addition, each instructor is in his or her office during a publicly-announced schedule of six hours a week for students to ask questions and clear doubts, either by visiting personally or through the University's *virtual campus* facilities.

3 Methodology in brief

3.1 Classroom work

Classroom work is organized around an *activity program*, a sequence of *activities* designed for students to learn the key concepts and basic techniques of the subject. Activities basically pose open problems that have to be tackled by the students, before the *theory* is explained. Here's an example of an introductory activity:

> Ambiguity is an essential feature of natural languages. Could you write up a formal definition of ambiguity? Why do you think human language is ambiguous? Why does ambiguity make machine translation difficult?

which is followed by a more detailed activity in which students are given a set of carefully chosen ambiguous sentences (with slight hints to clarify the various interpretations) and are asked to use those sentences and other examples they may come up with to design a linguistically motivated classification of ambiguity types.

Students organize themselves in groups of three (stable, if possible, for the duration of the course) to perform the activities; groups designate a spokesperson for each session. After working individually (sometimes at home, before the session) and in groups, a discussion takes place in the classroom, followed by a "classical" lecture segment where classroom work is integrated with the explanation given by the instructor. For example, after the ambiguity activities described above, the instructor introduces the *principle of semantic compositionality* (Radford et al. 1999) and therefore classifies ambiguities as *lexical* (the sentence has more than one interpretation because one or more words do: *polysemy*, *homography*, *anaphors*), *structural* (the sentence has more than one interpretation because it has more than one possible parse tree: *adjunction* and *coordination* ambiguities, ambiguity due to *Wh-movement*), or *mixed* (when both things happen simultaneously).

This way of organizing classroom work (based on a proposal used to teach natural sciences, Gil-Pérez and Carrascosa-Alis 1994) gives students the opportunity to analyse each problem and even advance parts of the solution, and prepares them to receive the solution when the teacher explains it after the discussion. But it also gives the instructor very valuable information on what students already know which he or she can use to *anchor* (Clement et al. 1989) the explanation of new, sometimes rather complex, concepts.

After the session, sudents are expected to make a synthesis between their individual and work group, the classroom discussion, and the explanations by the instructor, using the recommended literature and the office hours of instructors. The use of office hours to clear as soon as possible any doubts is strongly encouraged by instructors to avoid pre-exam *indigestion*.

3.2 Laboratory work

In each laboratory session, a computer assignment is proposed which has to be performed either individually or in pairs. For example, in unit 7 (lab session L_6) students are asked to analyse what does a given commercial MT system do besides simply substituting words, by first forcing the system to translate in isolation (e.g., each one in a paragraph) the words of a set of sentences and then translating the whole sentences (Pérez-Ortiz and Forcada 2001). In another assignment (lab session L_7), students feed a set of increasingly complex noun phrases designed by the instructors into an MT system to incrementally infer the word-reordering rules used by explaining the resulting correct or incorrect translations (Forcada 2000). The following section gives brief descriptions of the remaining laboratory assignments.

4 Syllabus

The current design of the syllabus started in 1995, before proposals like LETRAC (Badia et al. 1999) or surveys about the teaching of these matters (Balkan et al. 1997) were available. I basically interpreted the official description of the subject and made a quick survey of what other universities in Spain were doing (according to their webpages). After eight years of redesign, the course was eventually structured in 10 units or *blocks* ($\mathbf{B}_1 \ldots \mathbf{B}_{10}$), which will be briefly described and commented upon in this section. Classroom sessions are denoted $C_1 \ldots C_{19}$, lab sessions are denoted $L_1 \ldots L_{11}$; sessions C_{19} and L_{11} are spare sessions for doubt clearing, finishing laboratory assignments, or just to make up for a session which might have been postponed or cancelled (strikes, torrential rain).

4.1 The ten "blocks" in brief

Block: $\mathbf{B}_1$: What are we going to study?

Objective: Knowing the ways in which computers may be applied to translation; recognizing which parts of the translation task can be automated and which ones cannot; understanding the concept of *machine translation*; being able to distinguish the two main kinds of *computer-assisted translation*: *human-aided machine translation* and *machine-aided human translation*; being able to enumerate and describe computer tools useful for translation.

Classroom sessions: C_1 (week 1).

Lab sessions: None.

Block B_2: Computers and programs.

Objective: Acquiring basic concepts about how personal computers work, to improve their practical application and the understanding of their applications to translation: hardware and software; memory units (b, B, kB, MB, GB); RAM, magnetic and optical media; files, directories and directory structure; computer programs and instructions; CPUs, frequency, speed, and instruction sets; operating systems.

Classroom sessions: $C_2 - C_4$ (weeks 2 and 3).

Lab sessions: L_1 (week 3: analysing the hardware characteristics of the PC in the lab; creating and modifying a directory structure on a diskette).

Block B_3: Internet basics.

Objective: Acquiring basic concepts about the internet and about its application to the translation task: computer networks, internet as a network of networks, internet services of interest to translators (lexical databases, dictionaries, encyclopedia, texts, bitexts), URLs, IPs, names and domains, hardware and software needed for home and office access to the internet, the use of search engines to choose among various possible translations.

Classroom sessions: C_5 (week 3).

Lab sessions: L_2 and L_3 (weeks 4 and 5: searching for translations with Google; basics of HTML; building a webpage from a template and publishing it).

Block B_4: Texts and formats

Objective: Learn basic concepts about the storage, format, structuring, presentation, creation and manipulation of text documents: character encoding (ANSI, the ISO-8859 family, Windows CP-1252, Unicode, UTF-8), the use of format for presentation and structuring of content, and conflicts between the two objectives), common formats and their usage (RTF, HTML, PostScript, PDF, etc.); XML (well-formedness, validation using a DTD, separation of content and presentation through stylesheets); generating text through digitization and OCR or through speech recognition.

Classroom sessions: C_6 and C_7 (weeks 4 and 5).

Lab sessions: L_4 and L_5 (weeks 6 and 7: validating XML documents against a simple DTD; tagging a text according to a certain DTD and validating it).

Block B_5: Machine translation and applications

Objective: Learning what machine translation is and how it can be used in the real world despite its imperfections: assimilation and dissemination applications; human-aided machine translation (preediting, postediting, interaction, controlled languages); MT as a a component of communication systems (multilingual chat, translated web browsing); nonlinguistic requirements (speed, format preservation).

Classroom sessions: C_8 (week 6).

Lab sessions: none.

Block B_6: Ambiguity[1]

Objective: Identifying ambiguity as the main source of errors in machine translation, understanding the diversity of its mechanisms and learning to classify the ambiguity of a given sentence or statement: ambiguity through the principle of semantic compositionality; lexical ambiguity (homography, polysemy, anaphor, anaphor through empty categories), structural ambiguity (adjunction, coordination, movement of constituents), and mixed lexical-structural ambiguities; basics of ambiguity resolution in MT systems (statistical and rule-based methods).

Classroom sessions: C_9 and C_{10} (weeks 7 and 8).

Lab sessions: none.

[1] The fact that a complete block is devoted mainly to linguistic aspects may be surprising; however, we have found that our students arrive with severe deficiencies in basic linguistics, even after a mandatory subject called “Linguistics applied to translation”. Discussing ambiguity is an excellent way to review basic concepts needed to understand the ensuing blocks.

Block $\mathbf{B}_7$: How does machine translation work?

Objective: Knowing the main machine translation strategies and their implementation as distinct, consecutive phases or tasks; identifying these strategies by analysing the machine translation of real or synthetic texts (using a mechanical word-at-a-time, word-for-word translation called *model zero* as a reference model, Pérez-Ortiz and Forcada 2001): commercial systems as intuitive refinements over *model zero* (categorial homograph resolution, adding multiword expressions to dictionaries, rules for local reordering and agreement); the transfer architecture (analysis, transfer and generation; morphological, syntactic and semantic transfer, intermediate representations, linguistic information needed in each phase, modularity as an advantage); interlingua as null transfer and its advantages; inductive strategies (statistical MT, example-based MT).

Classroom sessions: $C_{11} - C_{14}$ (weeks 9 and 10).

Lab sessions: L_6 and L_7 (weeks 8 and 11: "machine translation is not word by word", Pérez-Ortiz and Forcada 2001, and "discovering reordering and agreement rules", Forcada 2000).[2]

Block $\mathbf{B}_8$: Machine translation evaluation

Objective: Learning to use knowledge about how MT systems work to evaluate them with an adequate technical level and well-founded criteria: aspects to be evaluated and their relative importance (quality, ease of use, extensibility, speed, memory usage), the difficulty of quality evaluation through postediting, the inadequacy of comparison with human translation, *predictive evaluation* after careful error diagnosis.

Classroom sessions: C_{15} (week 11).

Lab sessions: L_8 (week 12: evaluation and classification of MT errors in real texts).

Block $\mathbf{B}_9$: Lexical databases

Objective: Learning basic concepts about databases and applying them to lexical or terminological databases: tables, records, fields, index fields, ordering for faster search (dichotomic search), indexing for multiple orderings, updating; using lexical databases for specialized translation and terminological coherence; concept-based lexical databases and their fields (terms, definitions, subject, author, date, cross-references). Being able to design, create and maintain a lexical database using the suitable software.

Classroom sessions: C_{16} (week 12).

Lab sessions: L_9 (week 13: creating a small lexical database and performing searches over it).

Block $\mathbf{B}_{10}$: Translation memories

Objective: Understanding the importance of translation memories (TM) as an efficient solution to human translation with a high degree of repetitiveness: TMs as databases (translation units as records); bitext processing (segmentation rules, semiautomatic alignment, translation unit extraction); pre-translation (exact and approximate matches), advantages of TM-based translation work; the TMX standard.

Classroom sessions: $C_{17} - C_{18}$ (weeks 13 and 14).

Lab sessions: L_{10} (week 14: a taste of the complete TM cycle: alignment of a bitext followed by pre-translation and correction of a new text and TM updating).

4.2 Comparison to LETRAC

It is inevitable to compare, even if briefly this syllabus to the only detailed curriculum proposal available, LETRAC[3] (Badia et al. 1999):

- Forty-five hours forces us to make sacrifices and makes it impossible to include a great part of the LETRAC proposal. One could say that the weight given to Computers in Translation in Alacant does not match the importance given by LETRAC to the subject.

[2] The first assignment is programmed before actual work starts in the classroom and serves as a very nice introduction to the first activities of $\mathbf{B}_7$.

[3] It is surprising to see that translator associations like the International Translators Federation (FIT-ITF) and the American Translator Association (ATA) have no curriculum proposals of their own. There is another initiative (`http://www.lisa.org/leit/`) by LISA (Localization Industries Standards Association) but it is centered around *localization* and has therefore a narrower scope than the subject discussed here.

- The official description of Computers in Translation makes the study of MT mandatory, whereas LETRAC makes it optional (LETRAC, for example, gives more weight to translation memory).
- Desktop publishing (QuarkXPress, Framemaker, Pagemaker, Ventura, etc.), mandatory in LETRAC, is not taught in Alacant.
- LETRAC barely touches XML (maybe it was too early: it does mention SGML) and gives more weight to character encoding than to structure and presentation (in Alacant they have similar weight).
- The treatment of terminology is wide in LETRAC and very brief in the Alacant subject, which may be compensated for by mandatory subjects dealing with terminology totalling 10,5 credits.

5 Bibliography

Initially the course relied on the classical book by Hutchins and Somers (1992), articles such as (Hovy 1993), and a set of handouts which have eventually grown into a downloadable *textbook* or *lecture notes* (`http://www.dlsi.ua.es/˜mlf/iat/iat.pdf`) in Catalan.[4] Students are strongly encouraged by instructors to read other textbooks and materials (just to cite a few: Arnold 1993; Boitet 1996; citealpboitet96u2; Hutchins 2001; Hutchins 1996; Jacqmin 1993; Krauwer 1993; Lewis 1997; Nirenburg 1987; Sager 1993; Samuelson-Brown 1996; Somers and Rutzler 1996; Trujillo 1999; Vandooren 1993; Wojcik and Hoard 1996) in addition to class notes as a way of acquiring alternate views about the subjects and testing their comprehension of basic concepts.

6 Closing comments

It is hard to describe a whole course in half a dozen pages; this summary is presented in the hope that it may be helpful to instructors teaching in similar environments or facing similar time restrictions; in fact, all of the materials (in Catalan) are available to anyone interested (I could even consider translating selected materials into English on request).

Acknowledgements: I thank Andy Way for comments and suggestions on the manuscript.

References

Arnold, D. (1993). Sur la conception du transfert. In Bouillon, P. and Clas, A., editors, *La traductique*, pages 64–76. Presses Univ. Montréal, Montréal. English translation available: `http://clwww.essex.ac.uk/˜doug/papers/transfer.ps.gz`.

Badia, T., Freigang, K.-H., Haller, J., Horschmann, C., Huber, D., Maia, B., Reuther, U., and Schmidt, P. (1999). LETRAC curriculum modules. Available at `http://www.iai.uni-sb.de/letrac/home.html`.

Balkan, L., Arnold, D., and Sadler, L. (1997). Tools and techniques for machine translation teaching: A survey. Technical report, University of Essex. Available at `http://clwww.essex.ac.uk/group/projects/MTforTeaching/`.

Boitet, C. (1996). (human-aided) machine translation: a better future? In Cole, R., editor, *Survey of the State of the Art in Human Language Technology*. Available at `http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html`.

Clement, J., Brown, D., and Zietsman, A. (1989). Not all preconceptions are misconceptions: finding 'anchoring conceptions' for grounding instructions on students' intuitions. *International Journal of Science Education*, 11:554–565.

Forcada, M. (2000). Learning machine translation strategies using commercial systems: discovering word-reordering rules. In *MT 2000: Machine Translation and Multilingual Applications in the New Millennium*, pages 7.1–7.8, Exeter, UK.

[4] Catalan and Spanish are the official languages of the University of Alacant; the course is taught in Catalan.

Gil-Pérez, D. and Carrascosa-Alis, J. (1994). Bringing pupils' learning closer to a scientific construction of knowledge: A permanent feature in innovations in science teaching. *Science Education*, 78:301–315.

Hovy, E. (1993). How MT works. *Byte*, (enero):167–176.

Hutchins, J. (1996). Evaluation of machine translation and translation tools. In Cole, R., editor, *Survey of the State of the Art in Human Language Technology*. Available at `http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html`.

Hutchins, J. (2001). Machine translation over fifty years. *Histoire Epistémologie Langage*, 23(1):7–31.

Hutchins, W. and Somers, H. (1992). *An introduction to machine translation*. Academic Press. (Spanish translation available, published by Visor in 1995).

Jacqmin, L. (1993). Classification générale des systèmes de traduction automatique. In Bouillon, P. and Clas, A., editors, *La traductique*. Presses Univ. Montréal, Montréal.

Krauwer, S. (1993). Evaluation of MT systems: a programmatic view. *Machine Translation*, 8.

Lewis, D. (1997). MT evaluation: science or art? *Machine Translation Review*, 6:25–36.

Nirenburg, S. (1987). *Machine translation: Theoretical and methodological issues*. Cambridge University Press, Cambridge.

Pérez-Ortiz, J. A. and Forcada, M. L. (2001). Discovering machine translation strategies beyond word-for-word translation: a laboratory assignment. In Forcada, M., Pérez-Ortiz, J., and Lewis, D. R., editors, *MT Summit VIII, Proceedings of the Workshop on Teaching Machine Translation*.

Radford, A., Atkinson, M., Britain, D., Clahsen, H., and Spencer, A. (1999). *Linguistics: an introduction*. Cambridge Univ. Press, Cambridge.

Sager, J. C. (1993). *Language engineering and translation: consequences of automation*. Benjamins, Amsterdam.

Samuelson-Brown, G. (1996). New technology for translators. In Owens, R., editor, *The translator's handbook*. Aslib, London, 3rd edition.

Somers, H. and Rutzler, C. (1996). Machine translation. In Owens, R., editor, *The translator's handbook*. Aslib, London, 3rd edition.

Trujillo, A. (1999). *Translation Engines: Techniques for Machine Translation*. Springer, London.

Vandooren, F. (1993). Divergences de traduction et architectures de transfert. In Bouillon, P. and Clas, A., editors, *La traductique*. Presses Univ. Montréal, Montréal.

Wojcik, R. and Hoard, J. (1996). Controlled languages in industry. In Cole, R., editor, *Survey of the State of the Art in Human Language Technology*. Available at `http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html`.