The UPF Learner Translation Corpus as a resource for translator training

Anna Espunya

Abstract
The Learner Translation Corpus developed at the School of Translation and Interpreting of Pompeu Fabra University in Barcelona (LTC-UPF) is a web-searchable resource created for pedagogical and research purposes. It comprises a multiple translation corpus (English-Catalan) featuring automatic linguistic annotation and manual error annotation, complemented with an interface for monolingual or bilingual querying of the data. The corpus can be used to identify common errors in the students' work and to analyse their patterns of language use. It provides easy access to error samples and to multiple versions of the same source text sequence to be used as learning materials in various courses in the translator-training university curriculum.

Keywords: learner translation corpus; multiple translation corpus; LTC-UPF; error-annotation; English-Catalan translation; translator training

1. Introduction

The goal of general translation courses in a translation and interpreting degree is for students to develop and improve their competency in translation, in itself an amalgam of various competencies including communicative skills (reading and writing), expert use of documentary resources, wide-ranging knowledge and intercultural awareness, ample problem-solving strategies and the ability to transfer ideas from one language to another. Practice is a key component in the learning process, which means that students are commonly assigned translation tasks.

The errors identified by teachers grading those translations deserve systematic analysis as they reveal the difficulties encountered by the trainee translators. Current technologies provide reasonably user-friendly tools for the development of computer corpora to store student productions. Learner corpora have existed for two decades now, developed by scholars in the field of language learning for research and pedagogical purposes. Most follow in the footsteps of the *International Corpus of Learner English* (Granger 1993; Dagneaux, Granger and Meunier 2002; Granger, Dagneaux, Meunier and Paquot 2009) in the sense that the texts are monolingual original productions (generally argumentative essays) in the learners' second or foreign language.[1]

In 2008 the teachers involved in the General Translation courses of the Translation and Interpreting Degree at Pompeu Fabra University (Barcelona) set out to build the LTC-UPF, a learner translation corpus that could help improve curriculum coordination between them and the English language section.[2] The source language of the collected translations is English (as a foreign language), while the target language is Catalan (the students' first language). Students' translations were annotated for errors with a taxonomy drawn from translation pedagogy (see section 2.3).

---

[1] An exception to this approach is the PELCRA Learner Translation Corpus, developed at the English Department of the University of Łódz (Poland) by collecting translations (as opposed to original production) from Polish (native language) into English (Uzar 2002: 249, cited in Castagnoli 2009: 39) for the purposes of research and teaching of English as a foreign language. Its error annotation typology is from the EFL field as opposed to the translation pedagogy field.

[2] The project was funded by the Catalan government through the *Agència per a la Gestió d'Ajuts Universitaris i de Recerca* (AGAUR), code number 2008MQD00006. Anna Espunya belongs to the Grup CEDIT (Centre de Discurs i Traducció), funded by the Catalan government (code 2009SGR 771).

As the detailed survey in Castagnoli (2009: 37-43) reports, in the past fifteen years there has been a spate of projects collecting student translations in electronic format, from text banks to fully fledged annotated corpora, differing in aims, composition, annotation schemes and query capabilities.

One of the pioneers is the Student Translation Archive (see Bowker and Bennison, 2003), a text bank with a wealth of external data on the students' characteristics and the conditions of production. The texts are not tagged with linguistic information or information about errors. In Spain, the ENTRAD project is a collection of translated texts aligned with their source texts set up in 2005 at the University of Zaragoza (see Florén and Lorés, 2008). Errors are marked on each text through highlighting (using a colour code) and through graphical marks. Without proper error or linguistic annotation, they have to be downloaded as raw text and further annotated or searched with the existing corpus analysis tools.

The first corpus of learner translations including the tagging of translation errors proper is the *MeLLANGE Learner Translator Corpus* (LTC)[3], which comprises texts from various fields (legal, technical, administrative and journalistic) and translations of them done by both students doing translation degrees and professional translators (who provide reference translations for comparison with student versions). It covers several European languages (Catalan, English, French, German, Italian and Spanish) and was intended as a resource for translator training.

A similar resource is the Russian Translation Learner Corpus or RuTLC (Sosnina 2006), consisting of English STs and their translations into Russian as the native language. It has been manually annotated for errors with a typology of 40 items. It is used to identify the frequency and distribution of error types in order to detect the most common lexical, stylistic and grammatical errors in student translations in order to adapt and improve teaching practices and materials.

All these projects were designed with pedagogical aims, both theoretical, i.e. research into the acquisition of the translating competence and the role of training methodologies, and applied, i.e. developing materials for translator training. As for the research strand, the field is clearly in its infancy, judging by the scarcity of publications reporting results or even research programmes. Castagnoli's (2009) doctoral dissertation entitled "Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation", for which she developed the Multiple Italian Student Translation Corpus (MISTiC), is perhaps one of the few exceptions.[4] As for the applied strand, the field is perhaps more advanced, although the improvements in translator pedagogy achieved by teachers exploiting their resources are often not emphasised or fail to see the light of day in specialised publications. Bowker and Bennison (2002) include a section reporting several real applications of corpora created from their STA to translation training.

The aim of the present paper is to introduce the LTC-UPF English-Catalan corpus of translations produced by students (translation trainees) and to briefly present potential applications to teaching. The structure of the paper is the following: in section 2 I

[3] http://mellange.eila.jussieu.fr and http://corpus.leeds.ac.uk/mellange/ltc.html

[4] MISTiC is not annotated for errors, as it was compiled for a "traditional" corpus research study on explicitation. Castagnoli (2009: 84) reports a size of 59 source texts (30 English, 29 French) aligned in a one-to-many relation –with a few exceptions– to 482 Italian student translations. Multi-parallel and longitudinal analyses are possible, as there are several translations for each ST and each student contributed more than one translation.

present the resource, its composition, annotation and search interface. Section 3 is devoted to the potential applications of the corpus. Lastly I state my conclusions and suggest lines for future work.

2. The corpus: composition, processing steps and interface

2.1 Composition and student profile

At its present stage, the corpus contains 10 source texts and 194 translations written by the first- and third-year students doing the Translation and Interpreting degree between 2006 and 2011. Each translation is sentence-aligned with its source text. It is thus a *multiple* translation corpus (Castagnoli 2009: 5), since for each English source text there are multiple translations into Catalan (which may have been collected over several years), ranging from 2 to 42. More precisely, the raw number of versions for each source text is as follows: 1 text with 42 translations; 2 texts with 30-32 translations; 2 texts with 24-27 translations; 3 texts with 11-12 translations; 2 texts with 2 translations. The size of the English sub-corpus is 8,448 words (source texts range in length from 258 to 2,279 words). The size of the Catalan sub-corpus (i.e. the translation corpus) is 200,187 words.

Representativity, measured as the sample taken from the course population, is heterogeneous owing to organizational constraints. For some source texts, the corpus includes translations written by the whole student population on the course; for others, the sample consists of a third of the population, randomly selected (between 10 and 14 assignments); for two source texts, only two translations were arbitrarily selected by the instructor, based on pedagogical interest. Text types include non-fiction (informative/instructive and essay), as well as fiction (narrative, drama and screenplay). A list of the source texts can be found in the references section.

The corpus contains information about external text attributes which can be used as filters in corpus queries, namely, academic year, course, text type, type of translation, instructor and source text author. It does not provide data on each student's individual profile; therefore, as the corpus stands now it cannot be used for research into the effects of individual-related variables. Indirect data for the level of proficiency in English of groups of students are available from the institution either in the form of results for the entrance test (for academic years 2006-2007 and 2007-2008) or from the results of placement tests according to competence in the English language that students take at the beginning of the academic year (from 2008-2009 onwards). A majority of students rank between the B2 and the C1.1 levels of the Common European Framework of Reference.

2.2 Annotation and processing steps

The corpus is annotated linguistically and for translation error types. One of the novel contributions of our project is the procedure for error annotation. Rather than using an environment disconnected from the teaching activities, we recycle the output of the Markin grading software, which allows the teachers to upload documents and mark texts with error tags that they have previously defined.[5] The output format is compatible with xml tagging. Thus, the annotation for translation errors is a by-product of the manual error correction performed by the instructors in the context of their normal teaching activities.[6] Currently the translation corpus contains 4,238 error tags.

---

[5] Markin is a tool developed by Martin Holmes. Copyright by Martin Holmes and Creative Technology (Microdesign) Ltd. URL address < http://www.cict.co.uk/markin/index.php>

[6] We are indebted to Judith Domingo and Martí Quixal, at the time affiliated with UPF and now members of the Voice and Language team at Barcelona Media, for their assistance with the implementation of this recycling scheme.

The Markin output files are collected, anonymized, cleaned up and sentence-aligned with their English source texts. Once aligned, they are linguistically annotated. The metadata is entered before the text files are finally indexed.

Linguistic annotation is automatic only, as no manual disambiguation or tagging is performed on the output. For Catalan it comprises word, lemma, POS, fine morphological features and syntactic function and was conducted with the shallow parser CatCG (see Alsina et al. 2002), a tool developed for unrestricted written Catalan texts. The morphological tagger has a precision of 0.92% and a recall of 0.98% (Alsina et al. 2002: 310). Linguistic annotation for English comprises word, lemma and POS, and was conducted with TreeTagger.[7] The accuracy of TreeTagger, determined empirically on the Penn Treebank Corpus, is 0.9599 (Mihalcea 2003).

Because this is a corpus of student translations the accuracy of the Catalan tagger can be expected to drop in comparison with corpora of properly edited written original texts. Of the 2,511 unidentified tokens in the latest version of the Catalan monolingual subcorpus, 133 (5.29%) are instances of spelling mistakes and typographical errors, including missing spaces between words. The remainder can be traced to a variety of sources: the largest group is that of typographical marks, both in isolation and in contact with a word (m- and n-dashes, quotation marks, full stops and commas, etc.); otherwise, we find Catalan words absent from the CatCG lexicon: proper names, standard words that are perfectly acceptable as well as non-standard words or spellings. It must be noted that text types include drama and screenplay, where the simulation of speech and of the colloquial varieties is part of the learning process for students. Lastly, we find English words that remained in the text after translation (source text words that were preserved for various reasons).

2.3 Translation errors: a typology

The taxonomy of errors comprises 25 simple categories with no subdivisions. It is derived from standard references in translation pedagogy such as Delisle (1993). It was drafted by the team of lecturers on the translation courses, who have ample experience in correcting translations with the help of error typologies, with three criteria: it should provide informative feedback to students; it should be practical and hence motivating for teachers as annotators, i.e. it should ease the correction task rather than complicate it; and lastly it should be useful for corpus users. The resulting taxonomy covers a broad spectrum of phenomena compromising content, quality of linguistic and cultural expression and suitability for the purposes of the translation.

Content disparities between the source text and the target text are identified through the classes ‘Nonsense’, ‘Wrong Sense’ and ‘Inexact Sense’, in decreasing degrees of semantic disparity. The classes ‘Overexplicitation’ and ‘Underexplicitation’ refer to opposite degrees of explicitness of the translation in comparison with the source text, whereas the classes ‘Omission’ and ‘Addition’ refer to unmotivated variations in the content of the target text. Failure to adapt the target text to the target readership and to fulfil the translation brief is identified by the classes ‘Inadequate for the Purposes of the Translation’ and ‘Cultural Term’.

As far as linguistic expression is concerned, the taxonomy covers interference from the source language (here, English) with the classes ‘Borrowing’, ‘False Friend’, as well as

[7] TreeTagger is a tool for annotating text with part-of-speech and lemma information developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. URL address <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

'Frequency Calque'[8] and 'Literal Translation'. The label 'Hispanism' refers to interference from the second language spoken in the target community (Spanish in contact with Catalan, in Catalonia). 'Lexical Imprecision' involves wrong lexical choices, i.e. phenomena ranging from violations of semantic restrictions and collocation patterns to lexical/terminological imprecision. 'Grammar', 'Spelling', 'Punctuation', 'Coreference', 'Connectives', and 'Sound' cover the stylistic and grammatical quality of the target text.

Error categories have been defined to be as mutually exclusive as possible in order to minimize disagreement between teachers. Nevertheless, a certain degree of subjectivity in the appreciation of both the source and importance of the error is unavoidable. This must be taken into account when formulating hypotheses and designing the corpus experiments, but does not invalidate error annotation.

2.4 Query interface

The corpus is available in two different query configurations: as a parallel corpus (English Source Texts aligned at sentence level with their translations into Catalan) and as a monolingual corpus of multiple translations (Catalan Target Texts). Both configurations are hosted by the corpus query web-searchable platform IAC (Corpus Access Interface, see Domingo, Badia and Colominas 2010). IAC, which uses the IMS Open Corpus Workbench tools, allows queries of annotated linguistic information and error types; results can be filtered by metadata. The query platform also offers basic frequency searches on the Catalan component (relative and absolute frequencies).[9]

Query possibilities vary according to the configuration. The aligned English-Catalan parallel corpus configuration allows the user to perform keyword in context searches on the English component (lemma, word and POS), on the Catalan component (lemma, word, POS, morphosyntactic information, syntactic function and error type) and on both simultaneously. This configuration has great potential as a resource for teachers of English and Translation courses as it can help them explore specific aspects of their students' performance by combining queries for various types of information. For those interested in methodological research, searches restricted by certain external variables such as degree year are made possible by the metadata.

The Catalan monolingual configuration allows the extraction of frequency lists by word, lemma, POS, syntactic function and error type, and provides basic statistics.

2.5 Challenges related with error annotation

The first challenge is that each word or text sequence can be tagged for only one error (owing to a technical constraint on nested attributes by the version of CWB used for the interface). This forces teachers to set priorities for the error category they want tagged and introduces a pedagogical bias into the corpus. For example, if a word is both an instance of a spelling mistake and of a wrong sense, the bias will be towards form or content, respectively. Individual bias can be filtered through the metadata *instructor*.

Another challenge that arises is ambiguity as to the *locus* of the error. By way of example consider the case where an inflected verb form is tagged as an instance of an 'Inexact Sense'. The error might be pinned down to the tense (i.e. a discrepancy in the temporal information) or to the predication expressed by the base form (i.e. a discrepancy in the meaning of the predicate). To ensure that the predicate will be recognised by the linguistic analyser, the error tag must preserve the word's integrity. This has effects on the specificity of the information that can be obtained through

[8] Defined by Vázquez Ayora (1977: 102) as the usage of a given element or structure with a higher or lower frequency than is normal for original texts.

[9] Interested readers may contact the author to obtain permission of use.

quantitative corpus search. In the previous example, a search for <error type=Inexact Sense> combined with <POS=V> will return instances of wrong lexical choices together with wrong tense choices. A careful qualitative examination of the results will be necessary to classify them further. An alternative solution such as a richer taxonomy of errors must balance precision with size. The present taxonomy, with 25 errors, has proved practical for all the parties involved (teachers and students).

In this section I have described the corpus and the query interface, and have briefly discussed the challenges posed by error annotation. In the following one I consider various potential applications.

3. Teaching applications

This corpus has been designed as an aid for teachers, since they are the ones who can best put forward hypotheses, exploit the tools to test them, interpret the results and apply the newly gained knowledge at different stages of the learning process from curriculum planning to the design of materials. As an aid for planning, Bowker and Bennison (2002: 505) highlight the identification of patterns suggestive of "problem areas", which can help teachers "appropriately orient the curriculum or class discussions in order to focus on generally problematic issues". The concept of "problem area" is also fundamental in empirical approaches to research into the acquisition of translation competence (see, e.g., PACTE 2009).[10] The corpus provides an empirical approach to the quantitative definition of "problem areas" by means of distribution data, for instance of lemmas tagged as errors, since the volume of once-only occurrences will reveal the extent of individual variation as opposed to generalization. [11]

The user-friendliness of the search interface makes it suitable for student use. However, pedagogical concerns recommend a teacher-supervised use rather than completely free access. Learner corpora, and especially error-annotated ones, expose the best and the worst performances. While access to the full extent of individual variability in translation versions can enhance the critical skills of students, an excessive focus on errors and negative examples is potentially harmful for the affective dimension of learning.

The next section offers a brief sample of uses of frequency data, word lists and concordances.

3.1 Frequency data and word/ lemma lists

Data on the distribution of error categories can help relate errors in the translation courses to the weaknesses in the students' language skills such as reading comprehension (in the source language of the translation) or richness and correctness of expression (in the target language). Remedial action can then be taken in a course or across all subjects in the degree. For example, the fact that seven of the ten most frequently tagged-for error types concern formal aspects (Lexical Imprecision, Catalan Grammar, Punctuation, Literal Translation, Hispanism, Spelling and Syntax) reinforces the need for insisting on the quality of the target text. At the same time, the fact that

[10] The PACTE research group (Process of Acquisition of the Translation and Evaluation Competencies) defines "rich points" as "specific source-text segments that contained translation problems" (2009: 212). Rich points are key research constructs intended to isolate the hidden stimulus in an experiment.

[11] As in all quantitative corpus analysis, the statistical significance of results depends on corpus size. For a multiple translation corpus, size means not only raw word count but also the number of versions for each source text.

Wrong sense is the second most frequent category is a call for attention to the students' reading comprehension skills.[12]

Beyond translator training, corpus data can be applied in pedagogical and bilingual lexicography, as several of its error tags explicitly target lexical difficulties (see Espunya (2013) for an exploration of false friends and instances of lexical imprecision related to contrasts in semantic fields between the two languages).

3.2. Concordances

Concordances drawn from a learner translator corpus can be used to compare and contrast a range of solutions for the same source text sequence. Teachers can use them for class preparation to target specific difficulties, but error analysis could be incorporated as an exercise for students to enhance their critical reading and translation assessment skills. In this section I provide an example of how they can enhance coordination between the English grammar and the English into Catalan translation courses.

Consider the following excerpt, which was part of a text given to first-year students as a translation assignment (the boldface is mine):

> The landlady is expected to provide heating in the room used for private study and/or bedroom. As the cost of heating a home is considerable, and is a cause of great concern to most landladies, **unnecessary use of radiators and gas or electric fires should be avoided.** Students should remember, when they are going out, to turn off all lights and fires. [Cambridge Campus Accommodation Guide, Anglia Polytechnic, 1996-1997]

The clause beginning with "unnecessary" can be interpreted as a piece of advice (i.e. non-binding obligation) in which the modal 'should' is in harmony with the verb 'avoid' and the passive voice; or it can be interpreted as an indirect command, i.e. "do not use heating devices unnecessarily". Lexically speaking, 'avoid' appeals to the students to self-regulate their behaviour, rather than expressing prohibition openly. The query <Word = should> + <Word = be> + <Word = avoided> on the English component, provides a concordance list of 24 translated versions (from two different years).

As presented in table 1, half of the 24 concordances express a pragmatically acceptable correspondence for the modal expression (i.e. non-binding obligation, advice, necessity), while the other half comprises versions with stronger modal values (i.e. binding obligation, prohibition, denial of permission).

Modal values	N
A) Non-binding obligation	9
B) Necessity	2
C) Advice	1
D) Binding obligation	7
E) Prohibition	4
F) Denial of permission	1

Table 1. Distribution of modal values expressed by Catalan translations

[12] The raw numbers of tags for the top ten categories are the following: Lexical imprecision (n=671), Wrong sense (n=600), Catalan grammar (n=508), Inexact sense (n=404), Punctuation (n=380), Literal translation (n=325), Hispanism (n=222), Spelling (n=188), Syntax (n=187).

This distribution indicates a disparity in the students' interpretation of the semantic and pragmatic content of the sequence. The D-F solutions hint at an insufficient understanding of how the immediate linguistic context (the passive voice, the lexical choice 'avoid', etc.) as well as the text genre (guide) and situational context determine the strength of modality. This can be addressed in the English grammar courses.

Considered from the perspective of the translation class, a qualitative analysis will help the teacher and the students (if the concordances are used as class materials) to trace the potential source of error. For example, in Catalan, the target language, the choice of verb tense and mood can distinguish binding from non-binding obligation. The periphrasis of obligation *haver de* (the most obvious correspondence for English 'should') in the simple conditional mood (*s'haur**ia** d'evitar*) lacks binding force, whereas in the present or future tense of the indicative mood (*s'**ha** d'evitar, s'haur**à** d'evitar*), the reading that obtains is binding obligation. Students should be aware of this because Catalan is their first language, so the teacher might want to insist on the importance of the revision stage in order to validate their grammatical choices against the backdrop of the conventions of the text genre (here a guide as opposed to a set of rules).

Besides word-for-word translations, the concordances show acceptable translations involving changes in perspective, a strategy called *modulation* in the Comparative Stylistics literature (Vinay and Darbelnet 1977). For example, in one of the translations the negative perspective "unnecessary use" is turned into a positive one (*ús responsable* "responsible use"), i.e. "students need to use X responsibly". In this particular case, modulation does not compromise the illocutionary value of the clause. However, modulation may distort the message, for example when "should be avoided" is translated as "it is forbidden" and "it is not allowed" (*l'ús innecessari està prohibit* / *no està permès l'ús innecessari*). In both cases, the sentences read as house rules or even university regulations.

5. Conclusions

We have long known that student work and the corresponding teacher feedback provide unique data about the students' actual capacities and learning needs. The LTC-UPF provides an example of how corpus annotation and search tools can significantly improve the storage and systematic analysis of such data. In this paper we have focused on the potential benefits mainly for teachers, ranging from the identification of "problem areas" which may not necessarily agree with their preconceived ideas, to the provision of textual samples for task design in case-study and inductive analyses.

Because it is a multiple translation corpus, it has intrinsic value for research into inter-translator variability and the language of translation. As a corpus of student productions, it can be used to study the development of translation competence, particularly if compared with a corpus of translation experts. Neighbouring areas such as computer-aided error analysis and statistical machine translation could also benefit from the data it provides.

Building the LTC-UPF corpus has been a positive learning experience for the team. We have started to incorporate the corpus data into our didactic methodologies and to do research into specific problem areas. Future projects include: enlarging the corpus with new batches of translations at the end of each academic year, engaging our colleagues from other language combinations (French and German as source languages, and Spanish as the target language) so that we may study the performance of students across language combinations, and lastly, improving error annotation practices to reduce subjectivity.

References

Alsina, A., Badia, T., Boleda, G., Bott, S., Gil, Á., Quixal, M., & Valentín, O. (2002). CATCG: un sistema de análisis morfosintáctico para el catalán. *Procesamiento del Lenguaje Natural*, 29, 309-310.

Bowker, L., & Bennison, P. (2002). Translation Tracking System: A tool for managing translation archives. In *Proceedings of LREC 2002* (pp. 503-507). http://gandalf.aksis.uib.no/lrec2002/pdf/115.pdf. Accessed 12 March 2011.

Bowker, L., & Bennison, P. (2003). Student Translation Archive. Design, Development and Application. In F. Zanettin, S. Bernardini, & Stewart, D. (Eds.), *Corpora in Translator Education* (pp. 104-117). Manchester: St Jerome.

Castagnoli, S. (2009). *Regularities and variations in learner translations: a corpus-based study of conjunctive explicitation*. PhD Dissertation, University of Pisa.

Dagneaux, E., Granger, S. & Meunier, F. (2002) (Eds.), *International Corpus of Learner English*. UCL (Presses Universitaires de Louvain).

Delisle, J. (1993). *La traduction raisonnée: livre du maitre*. Ottawa: Presses de l'Université d'Ottawa.

Domingo, J., Badia, T., & Colominas, C. (2010). IAC: A dynamic corpus interface. In Xiao, R. (Ed.), *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies 2010 Conference (UCCTS2010). Edge Hill University, Ormskirk, 27-29 July 2010*. http://www.edgehill.ac.uk/uccts2010proceedings. Accessed 12 March 2011.

Espunya, A. (2013). Investigating lexical difficulties of learners in the error-annotated UPF learner translation corpus. In Granger, S. , Gilquin, G. & Meunier, F. (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use - Proceedings, 1, Louvain-la-Neuve: Presses universitaires de Louvain, 129-137.

Florén Serrano, C., & Lorés Sanz, R. (2008). The application of a parallel corpus (English-Spanish) to the teaching of translation (ENTRAD Project). In Muñoz-Calvo, M., Buesa-Gómez, C. & Ruiz-Moneva, M.A. (Eds.), *New Trends in Translation and Cultural Identity* (pp. 433-443). Newcastle upon Tyne: Cambridge Scholars Publishing.

Granger, S. (1993). International Corpus of Learner English. In J. Aarts, P. de Haan, and N. Oostdijk, (Eds.), *English Language Corpora: Design, Analysis and Exploitation* (57-71). Amsterdam: Rodopi.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (2009) *International Corpus of Learner English*, v2. Louvain-la-Neuve (Belgium): Presses universitaires de Louvain.

Mihalcea, Rada (2003). 'Performance Analysis of a Part of Speech Tagging Task'. In A. Gelbukh (Ed.), *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguists (CICLING), 2003, Mexico City, Mexico*. *CICLING 2003, LNCS 2588* (*pp.* 158–167). Berlin and Heidelberg: Springer.

PACTE (2009). Results of the Validation of the PACTE Translation Competence Model: Acceptability and Decision Making, *Across Languages and Cultures*, 10 (2), 207-230.

Sosnina, E. P. (2006). Development and application of Russian Translation Learner Corpus. In *Proceedings of Corpus Linguistics – 2006, St. Petersburg, Russia, 10-14 October 2006* (pp. 365-373).

Uzar R. (2002). "A Corpus Methodology for Analysing Translation". In Tagnin, S.E.O. (Ed.), *Cadernos de Tradução: Corpora e Tradução*, 1 (9), 235-263.

Vázquez-Ayora, G. (1977). *Introducción a la Traductología*. Washington: Georgetown University Press.

Vinay, J.-P. & Darbelnet, J. (1977). *Stylistique comparée du français et de l'anglais: méthode de traduction*. Paris: Didier (1990).

Source texts

Anglia Polytechnic: “Cambridge Campus Accommodation Guide”
Bogosian, Eric: “Our Gang”
Heaney, Seamus: “Crediting Poetry: The Nobel Lecture”
Institutional author: “Internet Safety: safe surfing tips for teens” (http://kidshealth.org/teen/safety/safebasics/internet_safety.html)
Publishing house: “Le Méridien Boston”, *Frommer’s 98 New England*.
Kelman, James: *How Late It Was, How Late*
Mishra, Pankaj: *The Romantics*
Rowling, J.K.: *Harry Potter and the Philosopher’s Stone*
Steig, W. and T. Elliot: *Shrek*. Screenplay.
The Economist: ‘Forests and How to Save Them’

Acknowledgements
The people involved in the project are, in alphabetical order, J. Ainaud, A. Espunya (coordinator), J. M. Fontana, M. Forcadell, M. González and D. Pujol. Ainaud, Pujol and Forcadell collaborated on the definition of the taxonomy; Espunya and Pujol are responsible for the collection and error annotation of translations. Fontana, González and Espunya are responsible for the exploitation of the corpus in the English language courses. The alignment and linguistic annotation of the texts was provided by J. Foraster and P. Giménez, on external contracts. Lastly, I would like to thank our colleague C. Colominas at UPF for her advice on technical issues regarding the construction of a parallel corpus.

I am very grateful to the anonymous reviewers for their very insightful comments and suggestions that have greatly improved this paper.