Designing a Learner Translator Corpus for Training Purposes

Sara Castagnoli (University of Bologna), Dragos Ciobanu (University of Leeds), Kerstin Kunz (University of Saarland), Natalie Kübler (University Paris Diderot), Alexandra Volanschi (University Paris Diderot)

Abstract

This paper presents the development of the MeLLANGE corpus, a multilingual, aligned and annotated learner tranlsator corpus (LTC). Unlike other learner corpora, MeLLANGE focuses on translation-related rather than language acquisition issues. The corpus contains thus students performance into their mother tongues. Moreover, the intention is not to deliver a repository od students errors, but rather give both trainers and trainees the opportunity to identify possible translation problems, as well as possible solutions, in a multilingual, corpus-based environment. The information attached to the corpus is divide into metadata regarding the soruce text, the translator and the translation situation, lingusitic annotation, and erro-annotation according to an error typology developed in the project. We show examples of how the MeLLANGE corpus can be exploited in teaching translation from a data-driven point of view, and from a data-based point of view.

1. Introduction

This article sets out to present the development of a multilingual annotated Learner Translator Corpus (hereafter LTC) – a corpus whose core is composed of translations produced by trainee translators and whose primary purpose is to provide insights into the most significant characteristics of such texts in order to inform translation pedagogy. The LTC was designed in the framework of the EU-funded Leonardo da Vinci project called MeLLANGE (*Multilingual e-Learning in LANGuage Engineering*) which aims at devising a methodology for the collaborative creation of eLearning teaching content in the fields of translation and translation technology, producing corpus-based teaching materials and, more ambitiously, establishing a framework for a European Master in Translation Technologies.[1]

The MeLLANGE LTC collects student translations of selected texts in all the language pairs commonly taught in the translation departments of the project's partners. In addition to these, the corpus contains some reference translations, namely translations produced by professional translators. Detailed information on how source texts (STs) were selected, on available language pairs, as well as on the whole collection procedure is provided in §3 below. In order to enhance the analysis of the student translations, a subset of the corpus was annotated with metadata and linguistic information, as well as error categories from an error typology specifically designed for the MeLLANGE project (see §4 below). A query tool was developed to take into account the exceptional structure of the LTC and enable queries involving error, metadata and linguistic categories – examples follow in §5.

The article concludes by presenting some statistics derived from the first analyses carried out on the LTC. Besides illustrating how the data were used for the immediate purposes of the MeLLANGE project, suggestions are made as to how these data can inform and support the "customisation" of teaching programmes and the development and improvement of (traditional and eLearning) teaching materials.

[1] Further details on MeLLANGE are available on the project's official website: http://mellange.eila.jussieu.fr

2. Project background and current research involving learner translation corpora

The idea that collecting and studying the output of trainee translators can provide useful information for translation teaching and research was first put forward at the end of the 1990s in pioneering projects such as Bowker and Bennison's *Student Translation Archive* (STA, Bowker & Bennison 2003) and the PELCRA project (Uzar & Waliński 2001). These principally aimed at identifying common problems in student translations – i.e. difficulties and errors – in order to improve teaching contents and materials. Similar objectives have been stated in the framework of recent studies, such as Florén's ENTRAD (2006; this volume) and Sosnina's *Russian Translation Learner Corpus* (RuTLC, Sosnina 2006). Overall, research of this type can be seen as standing at the crossroads between the translation corpora and the learner corpora research traditions, with translation students representing a special type of language learners whose *interlanguage* is the focus of investigation. Error analysis appears to be the most frequent type of analysis conducted on these learner translation corpora, even though they might also enable research on successful translation strategies and correct retextualisations.[2]

The four above-mentioned corpora differ mainly in terms of the languages involved, the translation direction, and the techniques/technology employed for corpus creation and querying. Most studies focus on translations into the students' native language in order to research translation-related phenomena, with the exception of the the PELCRA corpus, which was conceived as a tool for foreign language teaching (Uzar & Waliński 2001), and therefore includes student output in the foreign language. Table 1 summarises the main features of the four corpora referred to in this §as they relate to the MeLLANGE *Learner Translator Corpus*.

Corpus	**SL(s)**	**TL(s)**	**TL status**	**Error annotation**	**Availability**
STA	FR, ES	EN	native	no	no
PELCRA LTC	PL	EN	foreign	yes	no
RuTLC	EN	RU	native	yes	no
ENTRAD	EN	ES	mixed	yes	online
MeLLANGE	DE, EN, FR, ES	CA, DE, EN, ES, IT, FR	native	yes	online

[2] While pursuing a rather different goal, a corpus of student translations was also designed by Popescu-Beliş et al. (2002) for the automatic evaluation of Machine Translation systems. The authors' starting point was that MT evaluation tools would benefit from the existence of a set of reference translations against which to compute the similarity/difference of a further candidate translation, and that a corpus of student translations could be employed in order to have access not only to presumably gold-standard translations – produced by professional translators – but also to lower-quality, and even "imperfect", translations. This corpus has also been reported to be potentially helpful in "extract[ing] statistics about the types of translations mistakes, and the correlation between the distribution of mistakes in a translation and the grade scored by that translation" (2002:18), which makes it resemble the objectives set by MeLLANGE and the other corpora previously mentioned. However, its stated purpose, design and application set it apart and for the purpose of the present discussion it will not be analysed further.

Table 1. Corpora details

Being the result of a large-scale European project assembling partners from 7 countries, the MeLLANGE LTC is unique in terms of language coverage and number of students involved. What further distinguishes it from the other corpora under examination is its composite structure: it includes reference professional translations alongside student translations, with all texts being aligned at context level (see §5 below). Compared to the only other corpus available online – ENTRAD – it enables complex queries – e.g. queries involving error types – and allows users to display different competing translations for a given sentence.

3. Building the LTC

The MeLLANGE LTC, which is being extended constantly, comprises originals of four different text types together with translations of these texts done by students and professional translators. The text types selected are: legal, technical, administrative, and journalistic.

This choice was motivated by the need to cover a spectrum of text types that professional translators handle most frequently, as well as to provide texts with various degrees of difficulty. The chosen materials were selected as being of a reasonable size to be put to pedagogic use – i.e. 350 words on average – and as available in at least all of the languages regularly used in the project partners' translation classes. Consequently, the legal text was offered for translation out of *da, de, el, en, es, fi, fr, it, nl* and *pt*, the technical and administrative ones out of *de, en, es, fr* and *it*, and finally the journalistic text out of *ca, de, en, es, fr* and *it*.

Following consultations with industry partners, in order to mirror real-life translation projects we provided translators with supporting information alongside the written material to be translated. This led to the building of *translation kits* containing a translation brief, the source text, and the longer passage from which the source text was extracted, which also contains additional background information. These translation kits were published on a collaborative content management system (CMS) – a Plone platform also used for project internal communication and collaboration purposes.

Students and professionals have translated under various conditions and using different methods: the translation tasks were either integrated into curricular translation courses offered at the partner institutions – as in-class or home assignments – or carried out as voluntary exercises. Translators could use various tools and resources –dictionaries, the Internet, corpora, translation memories, terminological databases, etc. Their translations were submitted to the MeLLANGE project through a customised upload mechanism implemented on the Plone platform, consisting of an HTML collection form linked to a MySQL database, which also recorded information related to the translation conditions mentioned above, as well as other metadata such as the duration of the translation task, the translator's native, second (and third) languages, and his/her university qualifications.

Before moving any further, we would like to stress that the MySQL database should not be confused with the MeLLANGE *Learner Translator Corpus* proper. While the former is an invaluable mechanism for storing translations and information about the author of the translation and the circumstances in which he/she produced it, the latter combines this information with tokenisation, lemmatisation, POS tagging and translation error annotation according to the MeLLANGE error categories (see §4), and also provides a complex and user-friendly query interface to all these data.

The MySQL database currently stores 152 translations into Italian, 74 into French, 62 into English, 49 into Catalan, 22 into Spanish, 15 into German, 2 into Romanian and 1 into Slovak, the overwhelming majority of which were produced by trainee translators and a small minority by professional translators. Both students and professionals were required to provide translations into their own native language because the purpose of the LTC is to provide insights into the characteristics of texts translated by trainees, and not into how well learners master a second (target) language.

The next step in building the LTC after receiving the students' and professional translators' contributions was storing metadata related to the translator and to the conditions under which the translation was done. Having the translations stored in a MySQL database enables the easy retrieval of metadata together with the content to be post-processed and optimised for the LTC query tool. Annotators wishing to enrich the LTC with either or both linguistic information such as part-of-speech (POS) and lemma tags, and translation error categories, can download subsections of the LTC from the MySQL database which meet criteria relating to the text type, source and target languages and institutional affiliation of the translator.

Several POS taggers and lemmatisers have been used to generate the linguistic level of annotation, and a specialised tool – MMAX – was adapted within the project partnership to enable translation tutors teaching at the partner universities and having native knowledge of the target language to create the MeLLANGE error annotation level (see §4 below).

Within the LTC, professional translations are used as reference translation solutions. This allows a direct comparison of translation equivalents with and without errors thanks to a functionality of the LTC query tool which displays relevant corresponding contexts in the source and target languages – including any reference and student alternative translations available (see §5 for more details).

4. Annotating the LTC

4.1. Linguistic Annotation

The linguistic annotation of the translations stored in the LTC – i.e. tokenisation, POS tagging and lemmatisation – was performed automatically at each partner institution using either publicly-available or locally-developed tools. In order to ensure the coherence between the linguistic and error annotations, tokenisation of the translations to be included in the corpus was performed prior to both types of analyses. This operation is essential since tokens are the minimal units to which errors may be associated, i.e. the minimal markable elements (also known as "markables"). Thus, the tokenised files were also used as input files for the error annotation tool (see §4.2).

Table 2 provides a summary of the tools used for each language and of the size of the tagsets used by each partner:

	Tagger	Tagset size
EN	Tree-tagger	49 tags (Tree-tagger EN tagset)
DE	TnT-tagger	54 tags (Stuttgart-Tübingen-Tagset)
FR	In-house probabilistic tagger	43 tags (selected from the 300 tags used in the Paris 7 "Le Monde Corpus")
IT	Tree-tagger	52 tags (Tree-tagger IT tagset)
ES	Connexor tagger	36 tags (Connexor Tagset for Spanish)
CA	Catcg (Constraint Grammar formalism)	380 tags

Table 2. Tools and tagsets used for linguistic annotation

As the table above suggests, the tagsets used (apart from the one used for annotating translations into Catalan) are of a similar size. Partners agreed that, given the specificity of each of the 6 languages involved, designing a common tagset would not be a realistic expectation. Nevertheless, partners agreed on a "reasonable" size that would be appropriate for the main purpose for which the LTC was built, namely concordancing involving translation errors made by trainee translators. This resulted in the simplification of existing tagsets – for instance, in the case of a highly inflected language such as French, the tagset used only distinguishes between *common* and *proper nouns*, instead of a richer variant which also distinguishes between *masculine*, *feminine*, *singular* and *plural nouns*. For the purposes of the LTC query mechanism – i.e. in order to enable users to search for parts of speech in a more user-friendly way than by studying the tagsets for each language – various tags were grouped under more generic categories such as *verb*, *adverb*, etc.

4.2. Error Annotation

While the lack of a common tagset could be easily bypassed by using more generic and transparent categories, the partnership was aware that allowing annotators to mark errors within translations using their own conventions would considerably complicate the coherence and uniformity of the annotation – and, consequently, the retrieval and analysis of error types. Before starting the process of collecting translations, therefore, the consortium researched and defined an error typology which would include all the categories of errors expected to be found in student translations in all the languages involved. The underlying assumption was that providing pre-defined error categories would enhance consistency across translations, language pairs and annotators, while being aware that this could not result in the complete absence of subjectivity in the interpretation and annotation of "erroneous" parts.

The survey of existing translation evaluation models, tools and schemes carried out by Secară (2005), which takes into account both academic and professional contexts, was used as a starting point to reflect on principles and useful error categories for our purposes. It must be noted, however, that contrary to the models which were considered, the MeLLANGE error typology is not meant to contribute to any evaluative process, the focus being on *describing* and studying specific translation phenomena rather than giving any quality judgment. Therefore, it does not provide for the encoding of the perceived "seriousness" or errors. Following the first annotation tests, the typology has undergone a process of intensive discussion, with modifications being suggested also by translation teachers at partner universities.

The result is a hierarchical scheme (Fig. 1) based on the fundamental distinction between content-related and language-related errors. These two main categories are further divided into subcategories, such as *SL Intrusion* or *Terminology and Lexis*, which in turn group more specific error types, such as *Too Literal* and *Inappropriate Collocation*. Each error type is marked by a code which is attached to erroneous words/phrases/sentences during the annotation process.

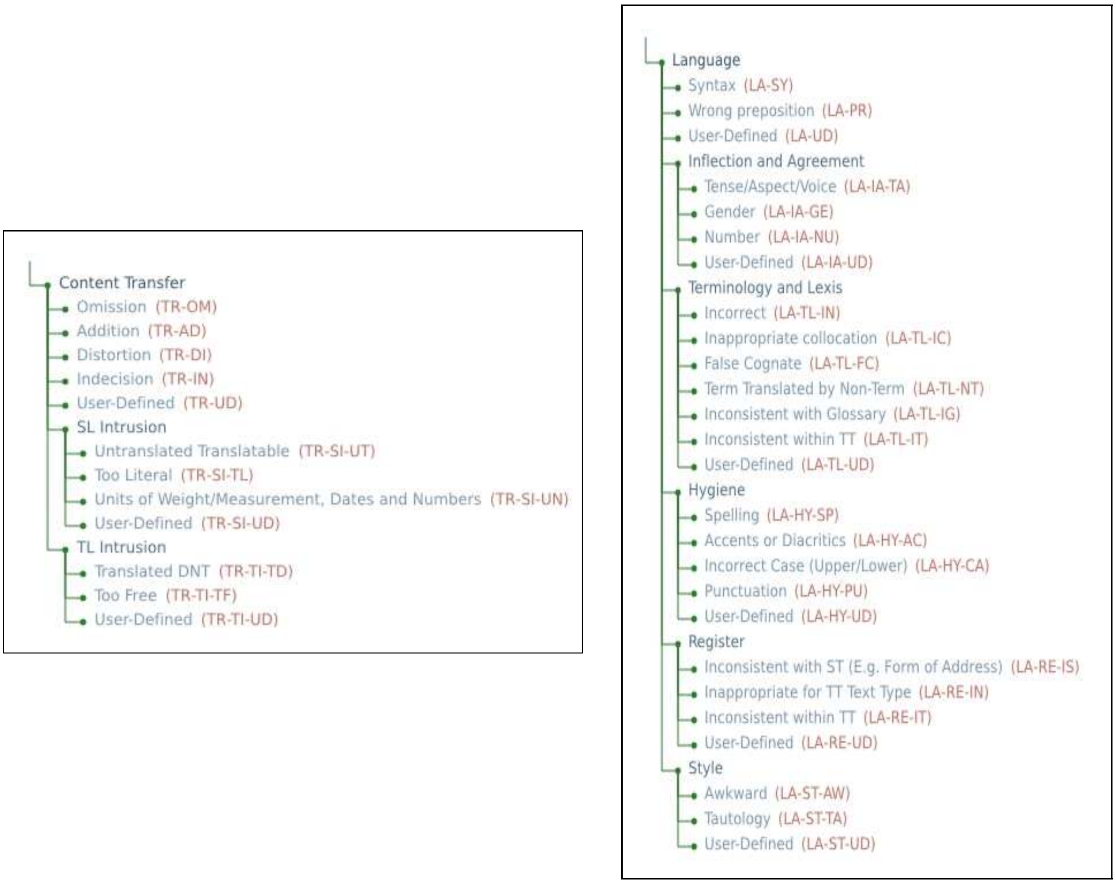


Fig. 1. MeLLANGE error annotation scheme

The taxonomy also includes *User-Defined* categories that annotators can resort to in case they find that the error they wish to mark does not belong to any of the other categories. It is also possible for annotators to assign more than one category to given text parts, as well as to provide explanatory notes and/or suggest correct solutions.

Error annotation is performed through a customised version of the manual annotation tool MMAX2, originally developed by Mark-Christoph Müller at the European Media Laboratory, Heidelberg (2006). The tool has an XML-based representation format and makes use of multi-layer stand-off annotation, which means that information about the annotation project is stored in different files. Our error annotation consists of four different levels: the Paragraph, Sentence, Content Transfer and Language Levels. While the Paragraph and Sentence levels are created automatically, the Content Transfer and Language levels were designed specifically for the annotation of the MeLLANGE LTC and reflect the different categories and subcategories of the error typology described above.

In order to create an annotation project, annotators do not need to be aware of such technical details, but simply have to save the translation in .txt UTF-8 format and drag-and-drop it on the “start_annotation” application. The annotation process is carried out using the MMAX4MeLLANGE graphical interface, which consists of several windows (see Fig. 2). The main window displays the text of the translation to be annotated. In this window annotators select the part(s) of the translated text which represent(s) either a Content Transfer or a Language error. The attribute window allows then for further specificity where annotators can select one or more of the pre-defined error categories or create a user-defined one, and leave appropriate feedback to students.

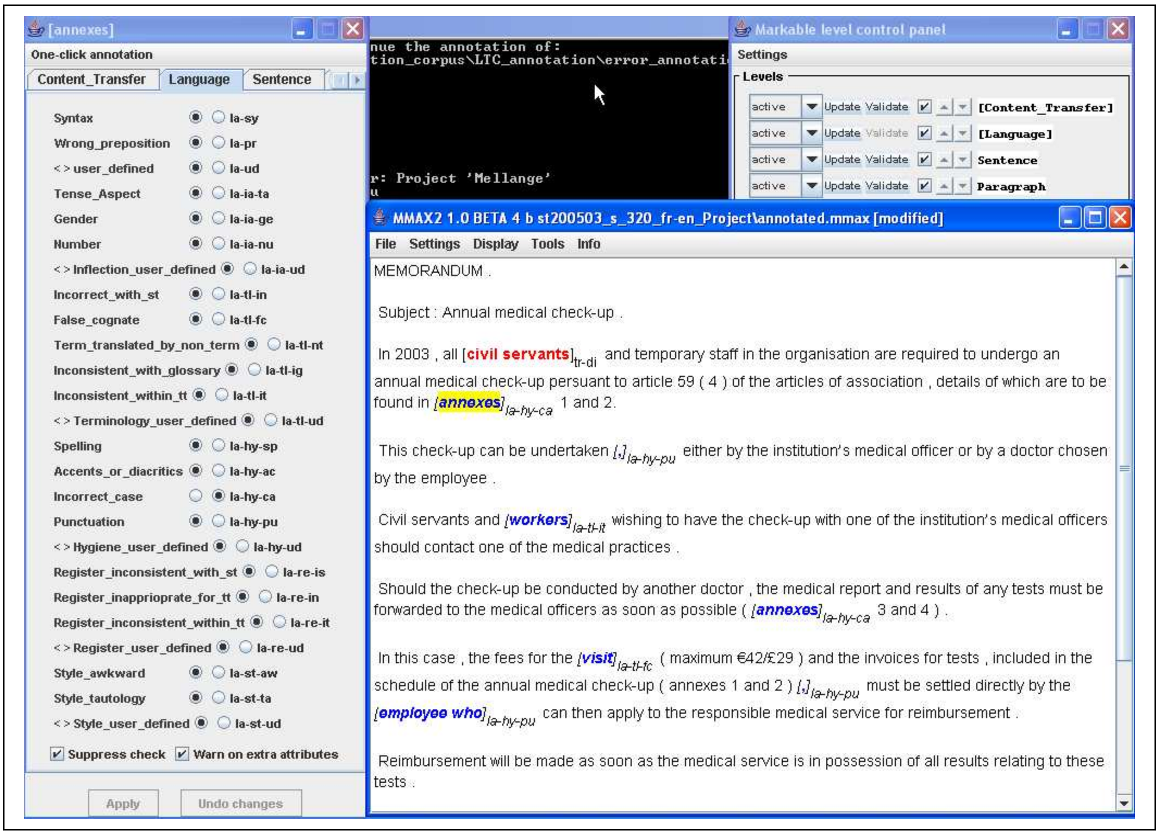


Fig. 2. The MMAX4MeLLANGE interface

5. Exploiting the LTC

Once translations from the MySQL database have been downloaded, tagged with linguistic information and also annotated using the MeLLANGE error typology, the final stage of post-processing is performed before they can be used by the LTC query tool. This stage consists of a normalisation of the output of linguistic taggers for the 6 languages involved in the project. As already illustrated in §4, several taggers and tagsets were used in the linguistic tagging process, with results such as those illustrated by Fig. 3.

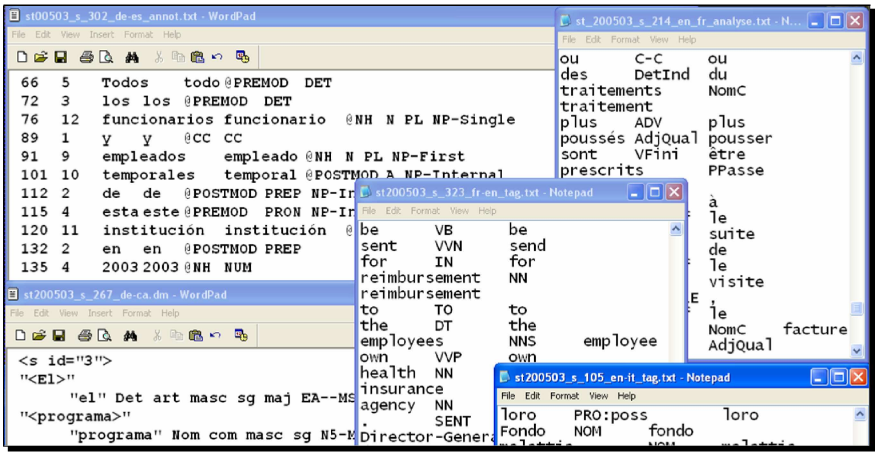
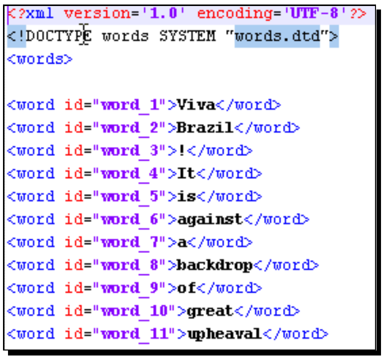


Fig. 3. Linguistic annotation output

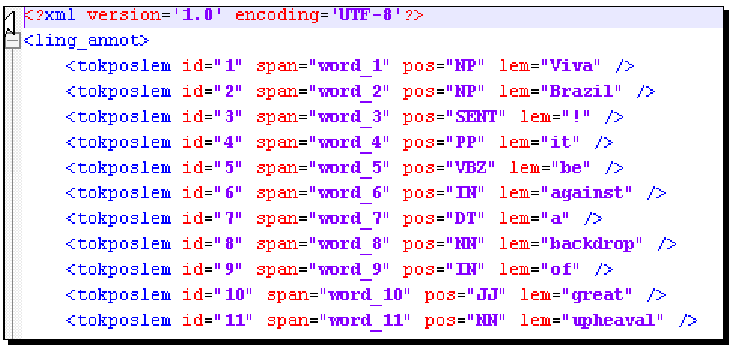
In order to integrate this information with the output of MMAX4MeLLANGE a certain degree of harmonisation was needed. Figures 4-6 illustrate the final formats in which the tokenised output, linguistic annotation, error annotation on the Language level (§4 has already highlighted that there is also a Content Transfer level) and metadata associated with a translation are respectively stored in the LTC. We chose to use stand-off annotation – i.e. storing each level of annotation in a separate XML file – because of the numerous advantages of this approach in terms of information management and maintenance.




```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE words SYSTEM "words.dtd">
<words>

<word id="word_1">Viva</word>
<word id="word_2">Brazil</word>
<word id="word_3">!</word>
<word id="word_4">It</word>
<word id="word_5">is</word>
<word id="word_6">against</word>
<word id="word_7">a</word>
<word id="word_8">backdrop</word>
<word id="word_9">of</word>
<word id="word_10">great</word>
<word id="word_11">upheaval</word>
```


Fig. 4. Extract from the tokenised version of a translation stored in the LTC




```
<?xml version='1.0' encoding='UTF-8'?>
<ling_annot>
    <tokposlem id="1" span="word_1" pos="NP" lem="Viva" />
    <tokposlem id="2" span="word_2" pos="NP" lem="Brazil" />
    <tokposlem id="3" span="word_3" pos="SENT" lem="!" />
    <tokposlem id="4" span="word_4" pos="PP" lem="it" />
    <tokposlem id="5" span="word_5" pos="VBZ" lem="be" />
    <tokposlem id="6" span="word_6" pos="IN" lem="against" />
    <tokposlem id="7" span="word_7" pos="DT" lem="a" />
    <tokposlem id="8" span="word_8" pos="NN" lem="backdrop" />
    <tokposlem id="9" span="word_9" pos="IN" lem="of" />
    <tokposlem id="10" span="word_10" pos="JJ" lem="great" />
    <tokposlem id="11" span="word_11" pos="NN" lem="upheaval" />
```

Fig. 5. Extract from the linguistic annotation of a translation stored in the LTC


```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml.org/NameSpaces/language">
<markable id="markable_18" span="word_58" la-st-aw_feedback="the"  style_awkward="la-st-aw" />
<markable id="markable_26" span="word_323" la-hy-sp_feedback="plummetted"  spelling="la-hy-sp" />
<markable id="markable_21" span="word_128" punctuation="la-hy-pu"  la-hy-pu_feedback="." />
<markable id="markable_24" span="word_200" la-re-in_feedback="unusual"
register_inapprioprate_for_tt="la-re-in" />
<markable id="markable_28" span="word_356" la-st-aw_feedback="reduced by a factor of 10? hit
extremely hard"  style_awkward="la-st-aw" />
<markable id="markable_25" span="word_291" la-hy-sp_feedback="employment"  spelling="la-hy-sp" />
<markable id="markable_23" span="word_175" la-st-aw_feedback="no real need for
sentence-initial and here, no real effect of style"  style_awkward="la-st-aw" />
</markables>
```


Fig. 6. Extract from the MeLLANGE error annotation (Language level) of a translation from the LTC


```xml
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="metadata.xsl"?>
<translation>
<user_info>
 <affiliation>Paris 7</affiliation>
<first_languages>en,pt</first_languages>
<second_languages>de,fr,it</second_languages>
<status>student</status>
<uni_level>postgraduate</uni_level>
<study_programme>Translation studies</study_programme>
<translation_studies_years>1</translation_studies_years>
<work_experience_as_translator>yes</work_experience_as_translator>
<work_years_as_translator>2</work_years_as_translator>
</user_info>
<text_info>
<source_text>ST200504</source_text>
<sl>fr</sl>
<tl>en</tl>
<years_using_sl>15</years_using_sl>
<years_using_tl>0</years_using_tl>
<reference_resources>yes</reference_resources>
<software_tools>yes</software_tools>
<local_peer_collaboration>no</local_peer_collaboration>
<remote_peer_collaboration>no</remote_peer_collaboration>
<expert_teacher_collaboration>no</expert_teacher_collaboration>
<assessed_translation>no</assessed_translation>
<had_time_limit>no</had_time_limit>
<time_spent>0</time_spent>
<ref_used_advice>no</ref_used_advice>
</text_info>
</translation>
```


Fig. 7. Extract from the metadata linked to a translation stored in the LTC

This modular structure of the LTC has enabled the building of a complex query tool, publicly available at http://corpus.leeds.ac.uk/mellange. The query tool enables users to explore resources produced in the MeLLANGE project, as well as one of its predecessors, the eCoLoRe project.[3]

[3] http://ecolore.leeds.ac.uk

While the interface to the eCoLoRe TMX corpus gives access to bilingual translation memories in 37 language pairs created from technical and financial translations, the MeLLANGE query interface also allows users to search the MeLLANGE LTC for examples of student errors, original and reference contexts, as well as student alternatives that meet various linguistic and/or metadata criteria (Fig. 8). A total of 232 annotated translations have been incorporated into the LTC (12 into *ca*, 13 into *de*, 54 into *en*, 49 into *es*, 48 into *fr*, and 56 into *it*).

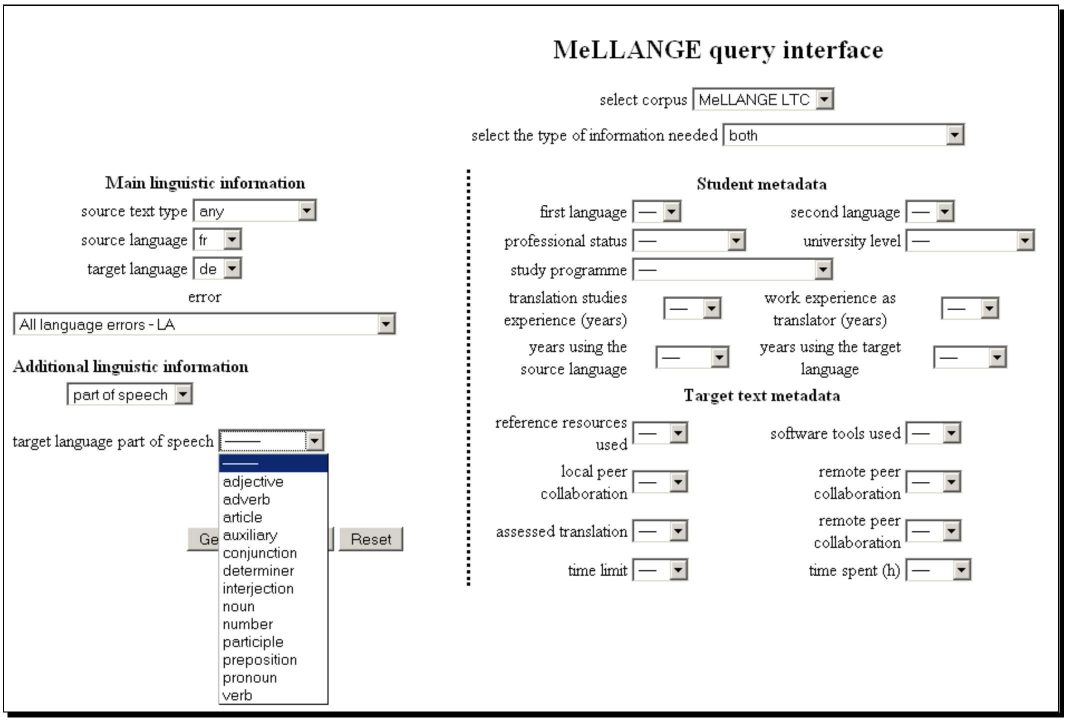


Fig. 8. Query interface to the MeLLANGE LTC

After using some of the available search criteria, the user is likely to be presented with a result such as the one illustrated by Fig. 9, which presents a typical answer to a query such as “display all sentences translated from *fr* into *en* which contain language errors as identified by the annotators using the MeLLANGE error typology”. First of all the user would be presented with the source context, then the target sentence which meets the criteria, followed by a reference context and alternative students translations if available.

1	Language	Context	Source
Original context	fr	Accès à l'emploi Tout ressortissant d'un État membre a le droit d'accéder à une activité salariée et de l'exercer sur le territoire d'un autre État membre, conformément à la réglementation nationale pertinente applicable aux travailleurs nationaux. Il bénéficie sur le territoire d'un autre État membre de la même priorité que les ressortissants de cet État dans l'accès aux emplois disponibles.	Full text
Target sentence	en	Every national of a Member State has the right to [undertake]$_{LA\text{-}TL\text{-}IC}$ gainful employment within the borders of any other Member State , according to the respective national [laws]$_{TR\text{-}DI}$ for [workers]$_{TR\text{-}OM}$.	Full text
Reference context	en	Focus on the employment Every national of a Member State has a right to work for a salary and to be employed on the territory of another Member State, as per the pertinent national regulations applicable to national workers. On the territory of another Member State, he/she is entitled to the same priority conditions as the nationals of this State to access to available jobs.	Full text
Alternative student context	en	Any citizen of a Member State has the right to access to employment within the territory of another Member State in respect of the relevant national [regulation]$_{LA\text{-}IA\text{-}NU}$ applicable to national workers . Wthin the territory of another Member State he is entitled to the same priority as national workers to access any available employment . He must be treated in the same way as national workers [looking for jobs]$_{LA\text{-}RE\text{-}IN}$ [through employment offices]$_{TR\text{-}DI}$.	Full text
Alternative student context	en	Any citizen of a Member State has the right to [access to]$_{LA\text{-}PR}$ employment within the territory of another Member State in respect of the relevant national regulation applicable to national workers . Within the territory of another Member State he is entitled to the same priority as national workers to access any available employment . He must be treated in the same way as national workers [looking for jobs]$_{LA\text{-}RE\text{-}IN}$ through employment offices .	Full text

Fig. 9. Query results from the MeLLANGE LTC

Given the short length of the texts to be translated and the limited probability of translators joining or splitting many sentences, for the time being the LTC is not aligned at sentence level. Instead, once a sentence produced by a trainee translator is identified as meeting the search criteria, its sentence id number (*X*) is extracted and a context of three sentences (*X-1*, *X*, *and X+1*) is returned from the source text, as well as the reference translation and student alternative translations.

Hovering with the mouse over tokens in the target sentence and alternative student contexts reveals POS and lemma information. The same procedure applied to the error codes reveals the feedback left by the annotators.

In the near future we plan to adapt this query tool to support the output of statistical information about the LTC, so that data such as those presented in Fig. 10 and Fig. 11 will be available just like the query results presented above.

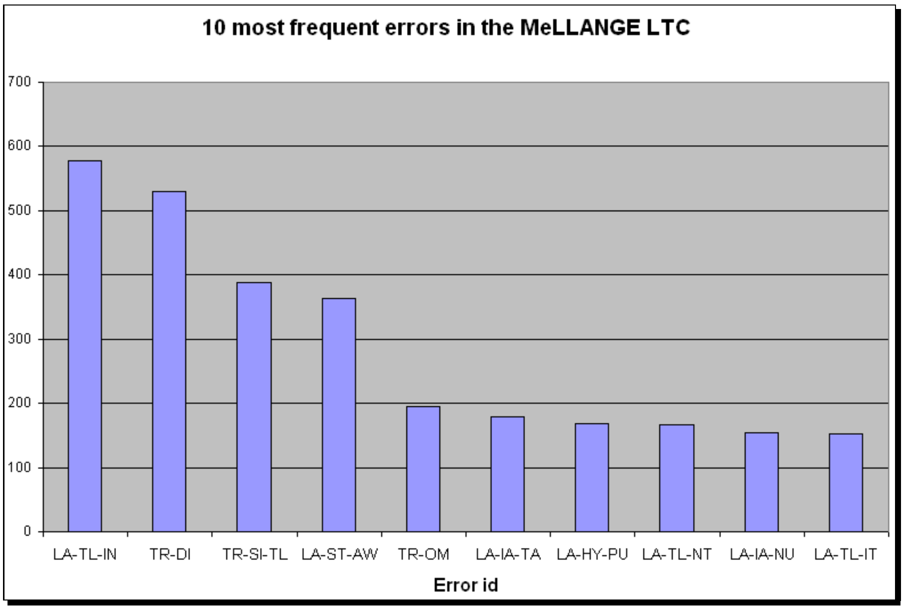


Fig. 10. 10 most frequent errors in the MeLLANGE LTC

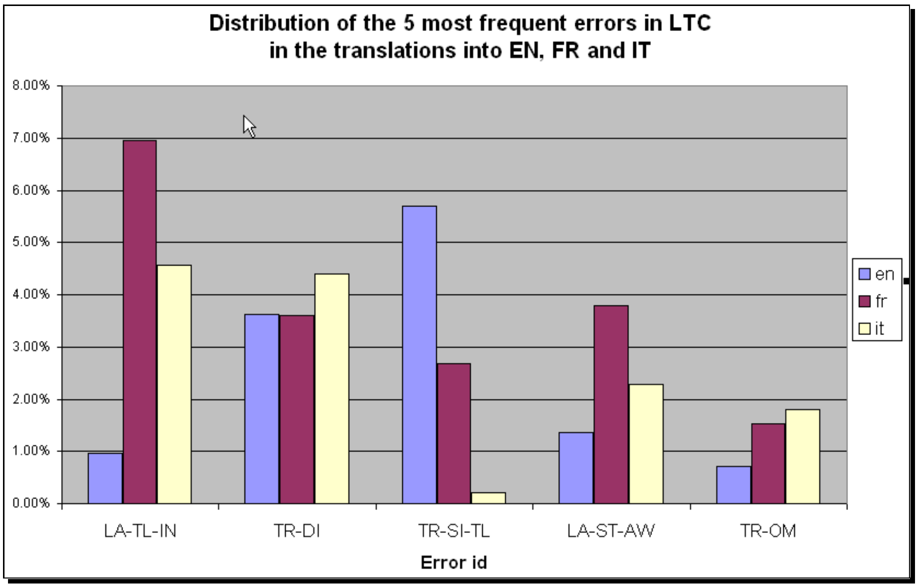


Fig. 11. Distribution of the 5 most frequent errors in the MeLLANGE LTC (translations into EN, FR and IT)

6. Possible applications of the LTC

Having introduced the two basic ways of consulting the corpus, namely by making parallel concordances and extracting statistics about error types, we will now provide practical examples of how the LTC data could be exploited in translation teaching and research, as well as of ways in which they have been used for the design of e-learning materials within MeLLANGE.

6.1 Using the LTC for didactic and research purposes

As far as the research into the quality and process of translation, as well as the teaching of translation, are concerned, the Learner Translator Corpus can – thanks to its different layers of annotation, associated to student metadata – enable researchers and teachers to look for specific types of errors in conjunction with particular word/lemma/pos sequences, and evaluate solutions found by translators of different backgrounds/levels of expertise to translation problems associated with particular text types. For instance, statistical analyses of the LTC could provide answers to questions such as:

- *what are the most common errors in the whole corpus* vs *in translations into a specific target language?* – which could highlight trends associated to the translation process itself or to specific language pairs (see Fig. 10-11 above);
- *are language errors equally frequent in students with different degrees of familiarity with the field of Translation Studies?* (see Appendix 1) – which would enable assessment of the influence of pedagogy on competence development;
- *do terminology errors tend to disappear in translations for which students had access to resources such as dictionaries, the web etc.?* – which could offer insights into the impact of reference resources on the quality of translations.

On the other hand, the concordancing function can display examples of, for instance:

- *content-transfer errors in EN-FR translations* (e.g. Fig. 12);
- *language errors made by undergraduate students,*

together with the ST sentence the errors are associated with, the reference translations available, as well as all alternative student translations for the same segment the error was found in (as discussed in §5 and shown in Fig. 9).

EN	He is entitled to the same social and tax benefits as **national workers**. A national of one Member State working in another **is entitled to** equal treatment in respect of the exercise of trade union rights, including the right to vote and to be eligible for the administration or management posts of a trade union.
FR	Il bénéficie des mêmes avantages sociaux et fiscaux que les [**travailleurs**]**TR-OM** . Un citoyen d'un État membre travaillant dans un autre État membre [[**peut**]**TR-TI-TF** **exercer** les mêmes droits syndicaux]**TR-TI-TF** [**que les travailleurs nationaux**]**TR-AD** , ce qui inclut le droit de vote et [d'éligibilité au sein de l'administration ou de la gestion]**TR-OM** d'une organisation syndicale .
REF	Il y bénéficie des mêmes avantages sociaux et fiscaux que les **travailleurs nationaux**. Le travailleur ressortissant d'un État membre occupé sur le territoire d'un autre État membre **bénéficie** de l'égalité de traitement en matière d'exercice des droits syndicaux, y compris le droit de vote et l'accès aux postes d'administration ou de direction d'une organisation syndicale.

Fig. 12. Content-transfer errors in an EN-FR student translation (edited) TR-OM = omission, TR-TI-TF = too free, TR-AD = addition.

While the statistical approach can help trainers identify the most common difficulties within a given group of learners, thus indicating areas of the learning curriculum where teaching is most needed, the concordancing functionality can provide teachers/researchers with authentic examples of the relevant error categories, which can then be used to produce either corpus-driven or corpus-based learning materials.

The term *corpus-driven* refers to learning content based/focusing on assessed difficulties – i.e. materials developed taking real student errors into account in order to better anticipate the problems that other trainees may face. This approach is based on the assumption that errors made by the students included in the corpus may be used to predict difficulties that other students translating into the project languages will have, and to build translation exercises accordingly.

By *corpus-based* materials, on the other hand, we mean materials which contain examples from the LTC. For instance, teachers might decide to tackle specific translation problems by providing students with concordance lines which show cases of those problems, asking them to spot the errors and think of alternative solutions. Or they might prepare exercises on translation revision using texts taken from the LTC, possibly exploiting the multiple alignment functionality to help student reflect on acceptable vs unacceptable variation. Some practical suggestions for developing corpus-based materials using LTC data are provided in Appendices 2 and 3.

6.2 Other scenarios

We would like to conclude by illustrating the usefulness of the LTC in two further scenarios: the teaching of other translation-related subjects, and autonomous learning. In the first case we will draw directly on the MeLLANGE experience of designing e-learning courses in translation-related disciplines, while the second is a suggestion whose applicability and usefulness still need to be assessed following empirical testing.

As was previously mentioned in the introduction to this paper, one of the aims of the MeLLANGE project was to produce e-learning teaching materials which would then serve as a basis for a European Master in Translation Technologies. One of the subjects that were identified by the partnership as essential for the translator curriculum was the use of corpora for translation. While most of the tools and subjects presented in other MeLLANGE courses (e.g. Translation Memory, Terminology, Localisation, Machine Translation) are likely to be well-known among translation students (and professionals) already, awareness of corpora and their usefulness for the practice of translation is unfortunately still scarce; therefore, we felt the need to first provide motivation for introducing these tools into the translation curriculum. To convince trainees and professionals that corpora could help them produce better texts, we built an exercise in which we asked them to revise a translation into English taken from the Learner Translator Corpus – focusing on the sentence below, which contains a lexical problem (underlined):

> *He shall be entitled to the same social and <u>fiscal benefits</u> as national workers.*
>
> Spanish ST: *Se beneficiará de las mismas ventajas sociales y fiscales que los trabajadores nacionales.*

The exercise consisted of various steps which aimed at showing – with the use of a specialised target language corpus – that the adjective *fiscal* is not usually employed to qualify *benefits*, and lead students to conclude that the expression *tax benefits* would have represented a much better choice in this context. Since the LTC contains authentic texts produced by trainee translators, we believe it can represent a very good source of data for any course dealing with language issues, as it offers examples of problems encountered by ‘real’ students in ‘real’ communicative situations.

Lastly, we also believe the LTC can be a powerful tool in situations of *blended* or *autonomous learning*, i.e. in cases where students are required/recommended to practice individually to supplement or replace traditional classroom teaching. Since learning to translate is not a matter of just “right” or “wrong” answers, translation is not a discipline which can be easily taught online. Very often students need guidance in order to understand errors and/or reflect on different translation options. However, we believe the LTC can contribute in a self-learning direction because:

- the reference translations produced by professionals represent “gold standard” versions against which students can assess the quality of their work;
- the translations of the same texts produced by other trainees give students the possibility to compare their translations with others in which errors have been marked and (often) corrected, thereby allowing them to see a range of translation solutions whose acceptability has been evaluated.

The possibility to query source texts, reference and trainee translations, aligned at a small-scale context level and enriched with several layers of annotations, resembles very closely the typical translation classroom situation, where several students are asked to suggest alternative solutions for an expression or sentence, which are then compared, discussed and evaluated by the teacher. Whether the Learner Translator Corpus can effectively be used as a tool for totally autonomous learning, however, still needs to be empirically tested.

References

Bowker L. & P. Bennison (2003) Student translation archive: design, development and application. In Zanettin F., S. Bernardini & D. Stewart (eds.) *Corpora in Translator Education*. Manchester: St Jerome, 103-117.

Florén C. (2006) ENTRAD, an English Spanish parallel corpus created for the teaching of translation. Paper presented at the *7th Teaching and Language Corpora conference (TALC 2006)*, Paris, 1-4 July 2006.

Müller C. (2006) Representing and Accessing Multi-Level Annotations in MMAX2. *Proceedings of the workshop on multi-dimensional markup in Natural Language Processing (NLPXML-2006),* EACL, Trento, April 2006.

Popescu-Beliş A., M. King & H. Benantar (2002) Towards a corpus of corrected human translations. In King M. (ed.) *Machine translation evaluation – human evaluators meet automated metrics, Workshop proceedings*, LREC 2002, 17-21.

Secară A. (2005) Translation Evaluation - a State of the Art Survey. *eCoLoRe/MeLLANGE Workshop Proceedings*, 39-44. (http://www.leeds.ac.uk/cts/research/publications/leeds-cts-2005-03-secara.pdf)

Sosnina E.P. (2006) Development and application of Russian Translation Learner Corpus. Presented at *Corpus Linguistics – 2006,* St. Petersburg, 10-14 October 2006.

Uzar R. & J. Waliński (2001) Analysing the fluency of translators. *International journal of corpus linguistics,* 6, 155-166.

Appendix 1 ***Using the LTC to explore research hypotheses***

A first type of activity that can be performed using the MeLLANGE LTC is researching how the quality of trainee translations can be influenced, among others, by time, resources, experience and level of training. Fig. 13, for instance, presents the difference between the work of a trainee translating into his/her mother tongue but with less than 2 years' experience in the field of Translation Studies. and that of another trainee with between 2 and 5 years of experience. Although we are guarded about generalising our results – mainly because we are currently expanding the coverage of the LTC – the initial results returned by the MeLLANGE query interface can already inform the drawing of research hypotheses.

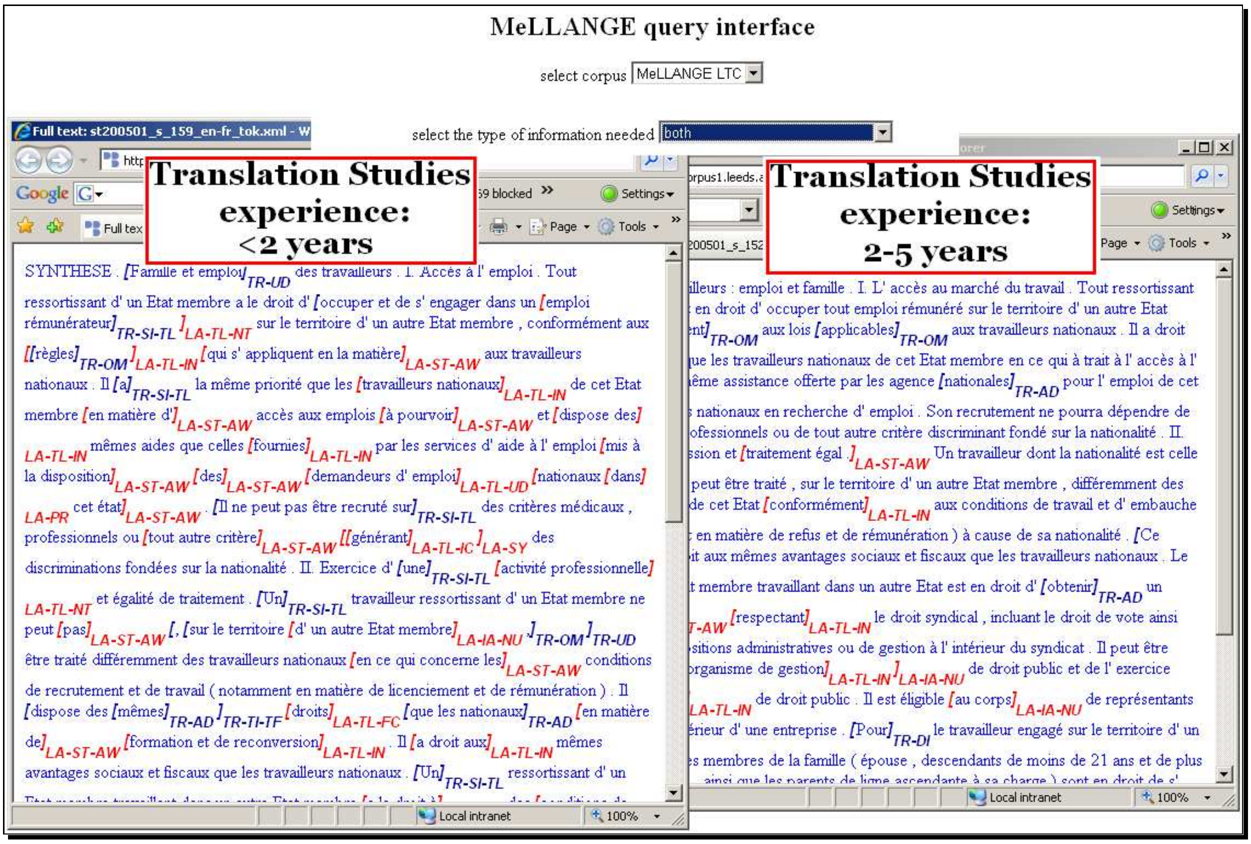


Fig. 13. Two translations produced by learners with different experience of Translation Studies

Appendix 2 *Translation error analysis using the LTC*

Error-annotated excerpts from the LTC can be shown to students to raise their awareness of specific error types. Students may be asked to discuss the errors from different perspectives – such as possible causes (e.g. ambiguities in the source text) and seriousness (i.e. the extent to which the source text meaning is affected) – and suggest corrections.

Example 1. Content transfer errors can distort the meaning of the source text (ST) to a lesser or greater degree. Fig. 14 shows an example in which the genuine meaning of the ST is slightly different in the target text (words in the ST that are concerned with the error are in boldface; content-transfer errors are tagged with a code starting with *TR*). The first error, tagged *TR-AD* (addition), indicates that an unnecessary word has been added to the target text: EN: *remuneration* – FR: *salaire **inégal***. Here, the adjective *inégal* is not necessary and alters the meaning of the ST, in which inequalities of treatment are implied, but not explicitly expressed. The second error *TR-OM* indicates that some parts of the ST have been omitted in the translation process: the equivalent of the English *in particular* does not appear in the French translation.

EN	Exercising an occupation and equal treatment. A worker who is a national of a Member State may not, in the territory of another Member State, be treated differently from national workers as regards working and employment conditions (dismissal and **remuneration in particular**) because of his nationality.
FR	[Un]*la-ud* travailleur d'un État membre [ne doit pas]*la-tl-it* être traité différemment [, sur le territoire d' un autre État membre ,]*la-sy* des travailleurs [de cet État]*la-tl-nt* , en matière de conditions d'emploi et de travail [(licenciement , [[salaire inégal]*tr-ad*]*la-tl-in*)]*tr-om* [en raison de sa nationalité]*la-st-aw* .
REF	Le travailleur ressortissant d'un État membre ne peut être traité différemment, sur le territoire des autres États membres, des travailleurs nationaux, en raison de sa nationalité, pour toutes conditions d'emploi et de travail (licenciement, **rémunération notamment**).

Fig. 14. "Less serious" content transfer errors. Edited LTC query interface results

Example 2. In the following example (Fig. 15), no word has been added nor omitted. Part of the meaning of the ST however, has disappeared in the translation process. Here, the complex preposition *on the grounds of* has been translated into *basés sur*, which is closer to *based on*. There is a loss in semantic substance in the target text, where *en raison de* would much more precisely yield the original meaning.

EN	His recruitment may not be dependent on medical, occupational or other criteria which discriminate **on the grounds of** nationality.
FR	Son [embauche]*la-tl-ig* [ne doit pas dépendre]*la-tl-it* de facteurs médicaux , professionnels ou [d' autres critères]*la-st-aw* discriminatoires [basés sur]*tr-ti-tf* sa nationalité .
REF	Son recrutement ne peut dépendre de critères médicaux, professionnels ou autres, discriminatoires **en raison de la** nationalité.

Fig. 15. "Too-free" content-transfer error. Edited LTC query interface results

Example 3. This example shows an occurrence of a serious content-transfer error, namely distortion. Distortion stands for a modification of the ST's meaning in the target text. Here, the distortion scope extends to the whole sentence. The prepositional phrase *under public law* is attached to the head noun *bodies* and further on in the text, to the head noun *office*. The French translation shows the prepositional phrase to be governed by the verb *excluded*, which completely changes the meaning of the sentence.

EN	He may be **excluded** from the management of **bodies** under public law and from the exercise of **an office under public law**.
FR	*[*Il*]*$_{la-hy-pu}$ peut être exclu de la gestion d' *[*organismes*]*$_{tr-om}$ et *[*être démis de ses fonctions *[*au nom du droit public*]*$_{tr-di}$ *]*$_{tr-ti-tf}$.
REF	Il peut être exclu de la participation à la gestion d'organismes de droit public et de l'exercice d'une fonction de droit public.

Fig. 16. Serious content transfer error

Appendix 3 ***Example of a revision exercise using the LTC***

It is also possible to use translations from the LTC without any annotation and ask students to spot errors and suggest revised translations (errors appear in boldface in Fig. 17 to facilitate reading, but they do not appear as such in the real exercise). As mentioned in §6.1, the multiple alignment functionality can also be used for triggering reflection on variation and translation acceptability, as students are allowed to analyse pros and cons of different translation solutions at the same time.

EN	Exercising an occupation and equal treatment A worker who is a national of a Member State may not, in the territory of another Member State, be treated differently from national workers as regards working and employment conditions (dismissal and remuneration in particular) because of his nationality.
FR	**Un** travailleur d'un État membre **ne doit pas** être traité différemment , **sur le territoire d'un autre État membre,** des travailleurs **de cet État**, en matière de conditions d'emploi et de travail (**licenciement**, **salaire inégal**) **en raison de sa nationalité** .
FR	Exercice d'une profession et **traitement égal**. Un travailleur dont la nationalité est celle d'un Etat membre ne peut être traité, sur le territoire d'un autre Etat membre, différemment des travailleurs nationaux de cet Etat **conformément** aux conditions de travail et d'embauche (plus particulièrement en matière de **refus** et de rémunération) à cause de sa nationalité .
FR	Exercice d'une profession et égalité de **rémunération**. Un travailleur, **citoyen** d'un Etat Membre **,** ne peut pas être traité différemment des travailleurs **du pays d'accueil** en ce qui concerne les conditions de travail et d' emploi (en particulier licenciement et rémunération) du fait de sa nationalité
FR	Exercice d'une profession et **traitement égal**. Un travailleur dont la nationalité est celle d'un Etat membre ne peut être traité, sur le territoire d'un autre Etat membre, différemment des travailleurs nationaux de cet Etat **conformément** aux conditions de travail et d'embauche (plus particulièrement en matière de refus et de rémunération) à cause de sa nationalité.
FR	Exercer une activité professionnelle et recevoir un **traitement égal**. **Un travailleur qui est** ressortissant d'un Etat membre ne peut pas, sur le territoire d'un autre Etat membre, être traité différemment des travailleurs nationaux **par rapport** aux conditions de travail et d'embauche (en particulier pour le licenciement et la rémunération) à cause de sa nationalité .

Fig. 17. A corpus-based exercise on translation revision

Students can then be shown the reference text so that they can compare it with their suggestions for revision. Fig. 18 is the reference text for the student translations shown above:

REF	Exercice de l'emploi et égalité de traitement. Le travailleur ressortissant d'un État membre ne peut être traité différemment, sur le territoire des autres États membres, des travailleurs nationaux, en raison de sa nationalité, pour toutes conditions d'emploi et de travail (licenciement, rémunération notamment).

Fig. 18. Reference French translation