Teaching Specialized Translation. Error-tagged Translation Learner Corpora

Jarmila Fictumova | Adam Obrusnik | Kristyna Stepankova
fictumov@phil.muni.cz | adam.obrusnik@gmail.com | 400362@mail.muni.cz
Masaryk University

Recibido: 15/12/2016 | Revisado: 16/02/2017 | Aceptado: 07/07/2017

Abstract

This paper describes the method used in teaching specialised translation in the English Language Translation Master's programme at Masaryk University. After a brief description of the courses, the focus shifts to translation learner corpora (TLC) compiled in the new Hypal interface, which can be integrated in Moodle. Student translations are automatically aligned (with possible adjustments), PoS (part-of-speech) tagged, and manually error-tagged. Personal student reports based on error statistics for individual translations to show students' progress throughout the term or during their studies in the four-semester programme can be easily generated. Using the data from the pilot run of the new software, the paper concludes with the first results of the research examining a learner corpus of translations from Czech into English.

Keywords: teaching translation; learner corpora; error-tagging; bilingual corpora; corpus-aided language learning; error statistics

Resumen

La enseñanza de la traducción especializada. Corpus textuales de traductores en formación con etiquetado de errores

En el presente trabajo se describe el método que se ha seguido para enseñar traducción especializada en el Máster de Traducción en Lengua Inglesa que se imparte en la Universidad de Masaryk. Tras una breve descripción de las asignaturas, nos centramos en corpus textuales de traductores en formación *(translation learner corpora, TLC)* recopilado en la nueva interfaz Hypal, que se puede incorporar en Moodle. Las traducciones realizadas por los alumnos se alinean de forma automática (con posibles modificaciones) y reciben un etiquetado gramatical y un etiquetado manual de errores. Es posible generar de manera sencilla informes sobre los alumnos con información estadística sobre errores en las traducciones individuales para mostrar su progreso durante el cuatrimestre o el programa completo. En función de los datos obtenidos en la prueba piloto del nuevo software, este trabajo presenta los primeros resultados del estudio a través de un corpus de traducciones de aprendices del checo al inglés.

Palabras clave: didáctica de la traducción; corpus textuales de traductores en formación; etiquetado de errores; corpus bilingües; aprendizaje de lenguas a través de corpus; estadística de errores

1. Introduction

Among the plethora of papers on improving translator training by introducing modern translation technologies and various software solutions published in recent years, this article stands out in its concern with one particular programme that has been developed and implemented in the course of several years of teaching specialized translation at the Department of English and American Studies of the Faculty of Arts, Masaryk University in Brno, Czech Republic. Two of the authors of this paper are graduates from the department. The aim of the teachers in the programme has been to provide the trainees with the necessary skills and competences to meet the requirements of the market and, at the same time, to put an emphasis on their autonomy in dealing with translation-related problems that they may face in future when they leave university. To this end, blended learning, online e-learning courses, the use of general and specialized corpora, and last but not least, Hypal, tailor-made new software, have been gradually implemented in the four-semester Master's English Language Translation programme that was accredited in 2008 by the Ministry of Education of the Czech Republic and The ECTS (European Credit Transfer System): Accreditation: 2008/02/21 – 2020/04/30. A part of this paper is devoted to the description of the programme and blended learning. The paper then deals with Hypal, the new software that is used for both teaching and research. The first part provides information about the technical aspects and various functionalities. Translation learner corpora are discussed in the next section, which also provides a detailed description of the Czech-English Learner Translation Corpus (CELTraC) compiled by Kristyna Stepankova. The corpus consists of student translations into English corrected by native speakers, and utilizes adjusted error typology.

2. English Language Translation-Programme Structure

The programme is built on the principles of EMT – The European Master's in Translation, a quality label for university translation programmes that meet agreed standards in education. The Master's programme includes both theoretical and practical aspects of translation, primarily English-Czech. The emphasis is placed on technical (non-literary) translation. A native-speaker knowledge of Czech is required. The core of the programme is a series of compulsory courses focused on translation theory, translation practice and Czech and English linguistics. The remainder of the programme consists of compulsory options dealing with specific types of translation and issues associated with them, such as subtitling, interpreting, terminology mining, legal and technical translation and others.

The rationale behind the teaching in the programme is based on the requirements specified in the EMT Translator Trainer Profile. As regards the general reference framework, the trainers (both practitioners and teaching staff) comply with the fundamental requirements, namely they have academic qualifications for university training (Master's and higher), relevant professional practice in translation, the ability to perform tasks assigned to students according to professional quality standards, appropriate teacher training and knowledge of translation studies and research relevant to particular courses. The competences to be met by the trainers are as follows: field competence, interpersonal competence, organizational competence, instructional, and assessment competence. All these requirements make it rather difficult to find suitable candidates. Our experience has been that it is well-advised to rely on former Masaryk University graduates working in various fields who are competent and willing to return to their *alma mater* to teach a course in their particular fields, although we are open to hiring experts from other universities and countries, provided they are fully qualified.

In the second semester each student is supposed to find a supervisor for their thesis. The topic areas are published and the students are expected to come up with their own suggestions, according to their fields of interest. Bachelor and Master theses can only be supervised by in-house faculty who are familiar with the resources available (for the list of translation trainees' theses please see Appendix 2).

Trainees are taught every week or every other week. The semester usually lasts twelve or thirteen weeks, with a Reading Week in the middle when trainees are advised to study on their own and do the recommended reading. The Reading Week usually includes a national holiday. Attendance in all courses is mandatory, one or two absences respectively are allowed per semester. As a rule, there is a written assignment for each week. Each course has its own e-learning support in ELF, prepared by the teacher, including resources and links to materials on the Internet. Teachers monitor trainees' work mostly through this system, setting deadlines for submitting written assignments. The system makes it possible to use discussion fora, peer assessment, team work, testing, etc. Trainees compile glossaries or upload materials for presentations in class, assess their peers' translations and their progress is constantly monitored. The assessment is based on trainees' continuous work throughout the semester, with particular emphasis on the quality of work, timely submissions of homework and activity in class. Feedback is provided both on individual basis and in class. Courses mostly finish with a final translation or a test, which is a part of the overall assessment. In some courses students are required to write essays in English, and in some other cases they are evaluated by means of their participation in a colloquium (in Practical and Technical Aspects of Translation).

In cooperation with the Faculty of Informatics, Masaryk University, employees and students have access to the latest technology in the field of corpus linguistics, namely Sketch Engine, for building corpora, both specialized and parallel. Following the trend and the suggestions that result from recent studies on corpus use by professional translators (Gallego-Hernández 2015; García Izquierdo & Conde 2012), we have concluded that the use of corpus managers in specialized translation work is and should definitely become an important part of translator training.

Many trainees in our translation programme have started using corpora while translating and do research as part of their final theses. Building specialized corpora to extract terminology in various domains in order to compile glossaries has been one of the main areas of co-operation with the experts from the Faculty of Informatics. The glossaries are often attached to trainees' diploma theses. Recent additions to the software (bilingual Word Sketch and bilingual term extraction from parallel corpora) have been tested by two diploma students (for the list of diploma theses please see Appendix 2). As suggested by Jaaskelainen & Mauranen (2005: 52), however, and based on their survey, "it seems that, in the case of freelance translators in particular, electronic tools are considered risky investments". The survey (2005: 53) "highlights the need for more cooperation in research and development between those who make the software products and those who end up using them". It can be said that this is the case of our trainees working with both Sketch Engine and Hypal and consulting the software developers when necessary.

In the classroom, one session usually lasts ninety minutes – i.e. two lessons. Due to current construction work within our campus, in some courses face-to-face lessons are held every other week. Having adopted a blended teaching approach, the trainers use the ELF e-learning system to supplement the teaching and to monitor regular students' weekly assignments. This teaching approach is frequently referred to as *mixed-mode, hybrid* or *blended learning.*

3. Blended learning

The last two decades have seen a growing trend towards e-learning and moving the course content online. This is true about both academia and the commercial sphere. To define the term *blended learning*, Margaret Driscoll (2002: 1), an IBM Global Services consultant, finds that the term refers to four different concepts, namely combining or mixing web-based technologies, various pedagogical approaches and instructional technologies with face-to-face teaching, or even combining instructional technology with actual job tasks "in order to create a harmonious effect of learning and working".

No matter how complex this definition may seem, it is true that our trainees experience various forms of blended learning during their studies, including the combination of learning with real-life jobs. They have been translating – under the supervision of their teachers – all English subtitles for the *Cinema Mundi International Film Festival* in Brno since 2010, as part of the *Subtitling* course. They participated in the *Translation for Heritage Promotion* Project, funded by Visegrad Funds. This involved cooperative training of translator trainees, producing six sets of English exhibition texts displayed in V4 museums and a heritage terms glossary. In another project [http://bit.ly/29xzLOr], namely *Terminology research for IATE*: *TermCoord, European Parliament* – Terminology Coordination Unit – trainees learned to build specialized glossaries for their subsequent inclusion in IATE in the following topic areas: *IT and Com-*

munications, Finance, Ecology (any subfield) and Economy. The project was meant to further develop these trainees' information mining and thematic competences.

Having most of the trainees' assignments and translations in electronic form has inspired trainers to exploit this material to its full potential and use it for compiling parallel corpora and learner corpora. To this end, new software was built and continues to be enhanced to serve the needs of translator training. As Fantinuoli and Zanettin (2015: 9) state:

> Corpus-based research critically depends on the availability of suitable tools and resources, and in order to cope properly with the challenges posed by increasingly complex and varied research settings, generally available data sources and out of the box software can be usefully complemented by tools tailored to the needs of specific research purposes. In this sense, a stronger tie between technical expertise and sound methodological practice may be key to exploring new directions in corpus-based translation studies.

The new tool, called Hypal (an acronym from Hybrid Parallel Text Aligner) should allow both teachers and trainees to achieve better results in their work. The main obstacle in compiling learner corpora is, however, data acquisition and pre-processing. Hypal was developed with the aim to make the pre-processing much easier and to integrate the data acquisition for learner corpora into the teaching process.

With regard to parallel corpora, the alignment of the texts is the most time-consuming task, unless an automatic pre-alignment algorithm is used. Hypal features such an algorithm, as well as a user-friendly tool for manual refinements of the automatic alignment (refer to section 4.2 for more details).

4. Introducing Hypal

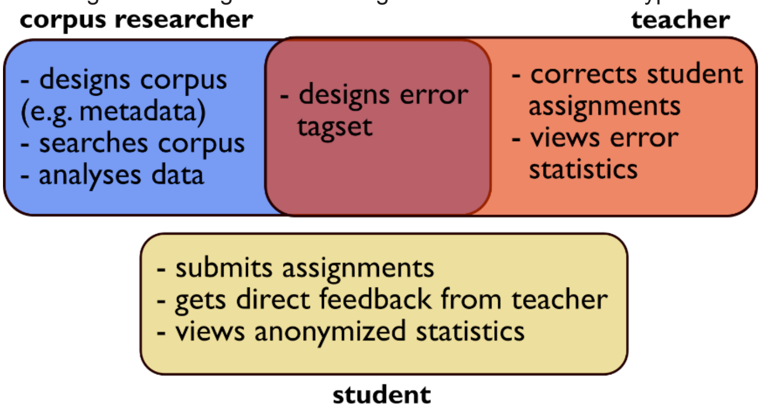
4.1. General structure and motivation

On the web page of Hypal [https://hypal.eu/hypal], its author states:

> Hypal is a user-friendly tool which is capable of **automatic and semi-automatic alignment** of parallel texts. The automatic alignment is primarily performed **based on statistics** and can be made more accurate by comparing **dictionary correspondences**. The alignment algorithm is analytical and therefore does not have to be trained. Furthermore, the alignment time scales linearly with the length of the parallel texts. More information about the alignment algorithm can be found in the author's Bachelor thesis [Obrusník 2013] [...] In addition, Hypal has extensive **error tagging capabilities**. Error tagging with a custom error tagset can be performed directly in the browser, **both on monolingual and parallel corpora**.

As suggested in the description above, corpora in Hypal are both monolingual and parallel. As far as learner corpora are concerned, the biggest challenge is data acquisition and error annotation of learner texts. Even though second language learners produce a large number of texts, which are subsequently corrected by language teachers, these texts are often available in printed form only or in a word processor format, which is not suitable for automated processing. Furthermore, individual teachers use different categories of errors, if any at all. Hypal aims to seamlessly integrate the acquisition of data for learner corpora into the teaching process by providing a web interface through which students submit their assignments and a web interface in which the teacher performs the error annotation using a set of pre-defined error categories and types. The error tagset, i.e. the set of error types sorted into categories, is in the ideal case designed by a corpus researcher working together with the language teachers, in order to make sure that the tagset is general enough but not too detailed as it would make the error annotation more time consuming for the teachers. The diagram in Figure 1 illustrates the distribution of user roles in Hypal.

Figure 1. A diagram illustrating different user roles in Hypal.



The first attempt at compiling a corpus included a parallel corpus to do research on professional translations. (Appendix 2, Knotková 2015) At the same time, teacher trainers started using Hypal for storing academic English papers. Error tagging was also introduced. (Appendix 2, Pokorná 2014.) It has to be pointed out that language teachers have to overcome an initial barrier to implementation of Hypal in their teaching as they have to learn to operate it first. To increase the motivation of teachers to use Hypal, they are provided with the option to view error statistics and analyse error reports of their students. This allows the teachers to identify the most frequent error types and incorrect phrases and they can integrate these into remedial exercises.

Later on, translator trainers' requirement for adjustments in Hypal was satisfied, and student translations started to be stored and subsequently error-tagged. This evolution was a considerably long process, with many different problems that had to be tackled. See section 7 for more information about the error tagging.

Before translation learner corpora, Hypal was also used for compilation of parallel corpora in specialized domains (legal, enviromental, etc.) in order to extract specialized terminology. This task was then performed in Sketch Engine. It is possible to export parallel corpora in *.tmx* format and import them in Sketch Engine like translation memories. If trainees work with specific translations and do not download the texts for a parallel corpus from the web, it is better to use Hypal because the alignment can be easily adjusted (see Figure 2) and is more accurate. Alignment in Hypal is more user-friendly than in some of the commercial computer-assisted translation (CAT) tools.

4.2. Alignment and PoS tagging

As already mentioned, parallel text alignment is a necessary pre-processing step when compiling parallel corpora, either error-annotated or not. The alignment algorithm in Hypal is capable of automatic alignment on sentence level, but it also has an intuitive interface allowing for subsequent manual refinements, described below. Practically, two language versions of a text can be split into sentences and aligned without further pre-processing but in reality, it is strongly desirable to perform part-of-speech (PoS) tagging of the texts prior to the alignment. When the texts are PoS-tagged, the searches in the resulting corpus are not limited to word form-based but the researcher can also form queries based on lemmas or parts-of-speech. Furthermore, with synthetic languages, the PoS tagging helps to improve the reliability of the automatic alignment, as the algorithm relies partially on finding lexical correspondences.

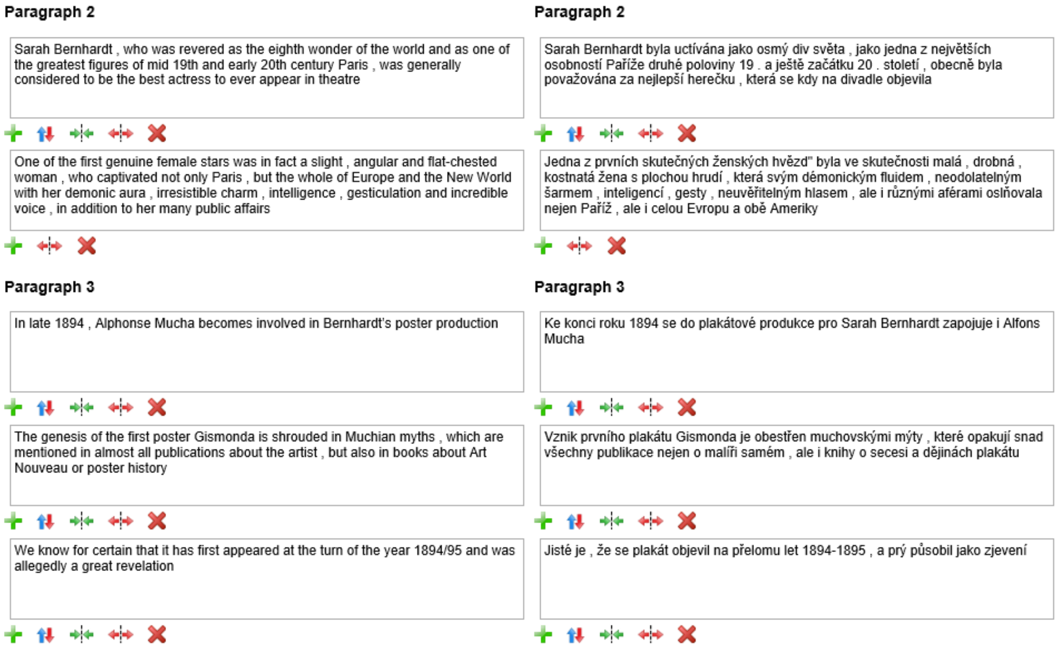
With regard to automatic alignment of parallel texts, two distinct approaches can be used, each bearing its own advantages and disadvantages. The common element of all parallel text alignment approaches is assigning a score to all possible pairs of sentences from the source language and target language texts, depending on their similarity and relative positions in the text. The final alignment, i.e. the mapping of sentences from source language to target language, is then chosen, so that maximum score is achieved. The scoring of a pair of sentences can be either performed by a trained artificial neural network (Tamura, Watanabe, Sumita 2014) or based on sentence lengths and lexical similarities in the two sentences, which is the approach that Hypal utilizes. The advantage of neural network-based algorithms over the algorithms relying on sentence lengths and lexical correspondences is their versatility, higher theoretical accuracy and their unchallenged performance when texts have to be aligned not on the sentence level but on the word level. On the other hand, neural network-based algorithms have to be trained first on a large body of diverse manually aligned texts, which is often not available. The automatic alignment algorithm in Hypal, on the other hand, relies only on the lengths of the sentences (defined below) and lexical correspondences between the sentences, and, as such, it works without requiring the time-consuming training phase.

In Hypal, the sentences are scored and aligned based on two main criteria. The first criterion is the length of the sentences, which can be seen as an extension of the Gale-Church scoring algorithm (Gale-Church 1993). However, unlike Gale-Church, Hypal defines the length of the sentence as the number of words in the sentence divided by the number of words in the whole text. This is expected to improve the performance of the algorithm, especially when aligning two language versions of the same text with one written in a synthetic language (such as Czech) and the other in an analytic language (e.g. English). Additionally, the sentence pairs are scored based on the number of lexical correspondences which are based on a bilingual dictionary. A more detailed mathematical description of the algorithm can be found in Obrusník (2013).

A few alignment algorithms, similar to the one employed in Hypal, have been previously implemented by other research groups, though they typically lack a user-friendly interface. The most similar algorithm is probably the Hunalign algorithm (Varga, Németh, Halácsy, Kornai, Trón, Nagy 2005) developed by the MOKK Centre at Budapest University of Technology. The main advantage of Hypal over Hunalign and most other sentence-level alignment algorithms is the capability to identify crossing alignments, e.g. when the sequence of sentences A B C in one language is arranged as A C B in the other.

Finally, in case the users do not agree with the automatic alignment, they can refine it directly in Hypal by using the manual alignment interface shown in Figure 2.

Figure 2. A screenshot of the manual alignment interface showing the buttons for adding a segment, exchanging two segments, merging two segments, splitting a segment into two and deleting the segment, respectively.



Paragraph 2	Paragraph 2
Sarah Bernhardt , who was revered as the eighth wonder of the world and as one of the greatest figures of mid 19th and early 20th century Paris , was generally considered to be the best actress to ever appear in theatre	Sarah Bernhardt byla uctívána jako osmý div světa , jako jedna z největších osobností Paříže druhé poloviny 19 . a ještě začátku 20 . století , obecně byla považována za nejlepší herečku , která se kdy na divadle objevila
One of the first genuine female stars was in fact a slight , angular and flat-chested woman , who captivated not only Paris , but the whole of Europe and the New World with her demonic aura , irresistible charm , intelligence , gesticulation and incredible voice , in addition to her many public affairs	Jedna z prvních skutečných ženských hvězd" byla ve skutečnosti malá , drobná , kostnatá žena s plochou hrudí , která svým démonickým fluidem , neodolatelným šarmem , inteligencí , gesty , neuvěřitelným hlasem , ale i různými aférami oslňovala nejen Paříž , ale i celou Evropu a obě Ameriky
Paragraph 3	**Paragraph 3**
In late 1894 , Alphonse Mucha becomes involved in Bernhardt's poster production	Ke konci roku 1894 se do plakátové produkce pro Sarah Bernhardt zapojuje i Alfons Mucha
The genesis of the first poster Gismonda is shrouded in Muchian myths , which are mentioned in almost all publications about the artist , but also in books about Art Nouveau or poster history	Vznik prvního plakátu Gismonda je obestřen muchovskými mýty , které opakují snad všechny publikace nejen o malíři samém , ale i knihy o secesi a dějinách plakátu
We know for certain that it has first appeared at the turn of the year 1894/95 and was allegedly a great revelation	Jisté je , že se plakát objevil na přelomu let 1894-1895 , a prý působil jako zjevení

The part-of-speech tagging and lemmatization in Hypal do not only extend the searching possibilities in the resulting corpus, but they also improve the performance of the alignment algorithm. Especially in synthetic languages, in which many words contain an inflectional morpheme, it is much easier to identify lexical correspondences at the level of lemmas than it is at the level of word forms. For part-of-speech tagging and lemmatization, Hypal relies on third-party tools which are also seamlessly integrated in the web-based user interface. Hypal currently integrates the MORCE tagger (Votrubec, Raab 2005), which is used for Czech and Slovak and TreeTagger, which supports English and a wide range of other languages. Hypal can work with parallel texts in any language pair, as long as the languages are supported by one of the part-of-speech taggers.

The following section is a brief overview of the development of translation learner corpora in general. It then goes on to introduce CELTraC, the first translation learner corpus compiled in Hypal.

5. Translation Learner Corpora (TLCs)

Translation learner corpora (or *learner translation corpora*, or even *learner translator corpora* – the terminology is not consistent yet) are parallel corpora compiled of source texts and their translations produced by "would-be translators in the process of acquiring requisite translating skills at a specific point/level of training" (Tiayon, 2004: 122). They play an increasingly important role in the fields of translation teaching and research, for they make it possible for the teachers to see how trainees learn to translate and what difficulties they encounter at a particular stage of translator training.

The idea that collecting and studying learner translations using a corpus approach can be of great help in the teaching of translation was first introduced at the end of the 1990s. *Student Translation Archive*, introduced by Lynne Bowker and Peter Bennison in 1997, and the *PELCRA* project, established by the University of Lodz in the same year, were among the first attempts to compile collections of electronically stored learner translations. These corpora, relatively small in size, primarily aimed at investigating trainees' *translationese* and identifying common problems and difficulties in learner translations (Bowker & Bennison: 2003; Uzar & Walinski: 2001).

These pioneers were then followed by the *Russian Translation Learner Corpus* (Sosnina: 2006) and the *ENTRAD* project (Florén, 2006). The first one – RuTLC – is reported to contain learner translations of technical texts. Similarly to the *PELCRA* corpus, it is also error-tagged, allowing automatic analysis of learner errors. The *ENTRAD* project is one of the three TLCs which are available online. It is a collection of about 45 English source texts and their translations into Spanish. The corpus is, however, not aligned at sentence level and thus cannot provide multiple concordances. The error typology is based on a colour code but it is not machine-readable (Florén: 2006).

Another remarkable achievement in translation learner corpus research was the compilation of the MeLLANGE LTC. The corpus was developed in the frame of the European-funded Leonardo da Vinci project called MeLLANGE (*Multilingual e-Learning in LANGuage Engineering)*, which was completed in 2007. The initial purpose of the project resided in providing translation teachers with corpus-based and corpus-driven teaching materials and researchers with a data bank that could be used for making comparative observations (Castagnoli et al. 2011). As far as language coverage is concerned, the MeLLANGE LTC is a unique resource – it contains over 400 learner translations in more than five languages (see Table 1). The corpus is compiled of originals of four different text types (journalistic, administrative, legal and technical) and their translations produced by learners as well as professional translators (to provide reference points for the trainees). In order to enhance the analysis of learner translations, the whole corpus was annotated with both with metadata about the translators, such as the translators' sex and level of experience, and various linguistic and error information (Kübler 2007). Thanks to its different layers of annotation, MeLLANGE allows researchers to search for specific error types in combination with particular word/lemma/part-of-speech sequences and to analyse target texts produced by translators with different levels of expertise (Castagnoli et al. 2011). Since the corpus is relatively big in size and easily available online, it represents a valuable source of data for universal deductions about translation trainees' performance.

Recent developments in the field of translation learner corpus research include RusLTC, KOPTE and UPF LTC. The RusLTC is a multiple learner translation corpus containing English and Russian source texts together with their translations produced by Russian translation trainees. The corpus is unique in terms of size – it currently contains more than 200 English source texts and approximately 1300 translations into Russian, and – since the corpus is bidirectional – over 40 Russian source texts and about 600 translations into English (Kunilovskaya et al. 2014). The dominant source text-types are essays and informational and educational texts. The corpus is error-tagged and, like the MeLLANGE LTC, annotated with various metadata about translators and translation situations. As far as availability is concerned, RusLTC is the third LTC available online after ENTRAD and MeLLANGE.

KOPTE is a translation learner corpus compiled as a part of a project launched by the Translation and Interpreting Department of Saarland University in 2009. The aim of the project was to develop corpus-based teaching materials that would assist both translation teachers and trainees. The corpus consists of French source texts (mainly newspaper articles) and their (nearly 1,000) translations into German produced by 58 German translation students. The whole corpus is annotated with part-of-speech and error information as well as metadata about the students (Wurm, 2013).

UPF LTC was introduced at the School of Translation and Interpreting of Pompeu Fabra University in Barcelona in 2008. The corpus is comprised of English-into-Catalan learner translations. In comparison with KOPTE, LTC-UPF is relatively small in size – it consists of ten source texts of various types (both non-fiction and fiction) and approximately 190 translations. The corpus features automatic linguistic annotation and manual error-tagging, and is complemented with an interface for querying of the data (Espunya 2013).

An overview of the main features of the above-mentioned corpora can be found in Table 1.

Currently there is a new corpus collection iniciative by Sylviane Granger and her team from The Centre for English Corpus Linguistics of Université catholique de Louvain, a project called MUST (Multilingual Student Translation) that started in the

Spring 2016. More details can be found in their website [http://www.learnercorpusassociation.org/must-multilingual-student-translation/].

Table 1. The main features of the TLCs.

Corpus	SL(s)	TL(s)	TL status	Error Annotation	Availability	Reference Professional Translations
STA	FR, ES	EN	native	no	no	no
PELCRA	PL	EN	foreign	yes	no	no
RuTLC	EN	RU	native	yes	no	no
ENTRAD	EN	ES	mixed	yes	online	no
MeLLANGE	DE, EN, FR, ES	CA, DE, EN, ES, IT, FR	native	yes	online	yes
RusLTC	EN, RU	EN, RU	mixed	yes	online	no
KOPTE	FR	DE	native	yes	no	no
LTC-UPF	EN	CA	native	yes	no	no

Our first translation learner corpus called Czech-English Learner Translation Corpus – CELTraC (Štěpánková, 2014), was recently compiled, composed of Czech into English student translations. Even though, as mentioned before, the main focus of the specialized translation programme is traditionally translation into the mother tongue, the trainees need to translate into English as well, due to the situation on the market. In a specialized course called *Special Topics in Translation: Translation into English* the trainees are supervised by two translation trainers – a Czech and an English native speaker. The CELTraC corpus currently consists of ten source texts and 341 translations produced mostly by Master's in Translation trainees in the above mentioned course, but also by Bachelor degree students from the *Becoming a Translator* course.

Table 2 provides detailed information about the amount of individual translations, the genres of the source texts, and the types of assignments (home or in-class). The first three translations are from the *Translation Approaches to the Humanities and the Media* course taught by an external translator trainer who is a freelance translator. The titles of the source texts are the names of the tasks in Hypal, as provided by the teachers.

Table 2. CELTraC: Current corpus data (23 March 2016).

Source Text	Genre	Author of Translation	Number of Translation	Type of Assignment
Sarah Bernhardt	exhibition brochure	MA	74	home
Lipník nad Bečvou	tourist brochure (history)	MA	77	home
Lomnice nad Popelkou	legend	MA	76	home
Pekáček	manual	MA	7	home
Letter	formal letter	MA	9	home
Článek *MUNI*	popular science (academic article)	MA	22	home
Co je překladová paměť	advertisement (translation memory)	BA	26	home
VK Final Translation: životní *prostředí*	popular science (environment)	MA	8	in-class
Břidlicový plyn	academic	BA	23	home
Břidlicový plyn 1	academic	MA	19	home

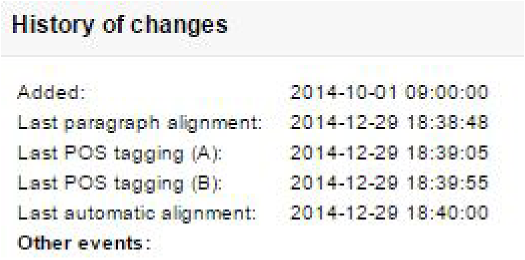
6. Translation error typologies

As explained earlier, Hypal now includes both monoligual and parallel learner corpora. The monolingual academic English corpora use their own error typologies, based on the nature of the purpose they serve – some of them are used in teacher training, others in linguistic research.

When the decision was made to create a translation learner corpus, it was decided to adopt the MeLLANGE error typology, since it seemed most appropriate for the task at that time. The translations to be analyzed and error-tagged in Hypal were translations into English, corrected by a native speaker of English. The manual error tagging was done by several people. It transpired that each of them understood the error categories in their own way, which resulted in having the same error marked in various ways, even though the corrections were by the same teacher. In the end it was clear that the typology needed to be adjusted. Šimon Javora based his Bachelor thesis on this topic and prepared a new error typology with definitions and examples of errors and their corrections, based on CELTraC error annotations. At the same time, a new functionality was added to Hypal. Figure 3 shows that the time of the upload of the text as well as all changes that were made to it are recorded.

In this way it is easier to track back the history of a text and, if necessary, make changes in error tagging. The new typology was then adjusted by Kristyna Stepankova (see the current version in the Appendix 1). It was applied to the old corpus (the tagging was transferred) and used for the new translations as well. To a certain degree the simplified error typology is still based on MeLLANGE, but at the same time, the colour-coding that has been used for correcting translations at the department before the introduction of an error typology, was taken over, namely that used for correcting translations in ELF (mostly translations into Czech). It will therefore be easier to transfer old corrected translations to Hypal and do the error-tagging using the same colours, only in a more detailed fashion.

Figure 3. History of changes.



History of changes

Added:	2014-10-01 09:00:00
Last paragraph alignment:	2014-12-29 18:38:48
Last POS tagging (A):	2014-12-29 18:39:05
Last POS tagging (B):	2014-12-29 18:39:55
Last automatic alignment:	2014-12-29 18:40:00
Other events:	

At this point another issue came to light. In the previous translations, teachers always highlighted very good translation solutions, which were then transferred to Hypal as *Positive feedback*. It is very important for translation trainees to see that one particular translation problem may have more than one good solution and compare these to their own translations. The role of praise and positive feedback cannot be underestimated by leaving these markings out. What, however, seems to be inevitable and obvious, is the fact that correct error statistics cannot be produced with *Positive feedback* (it is not an error). For this reason a filter was added, where *Positive feedback* can be shown separately and it is not included in error statistics. *Positive feedback* and each error have an abbreviation that appears in brackets on the left above the annotated word(s) (see Figure 5) and can be used for searches in the corpus. In this way all excellent solutions marked by the trainer can be shown to the trainees.

7. Corpus Compilation

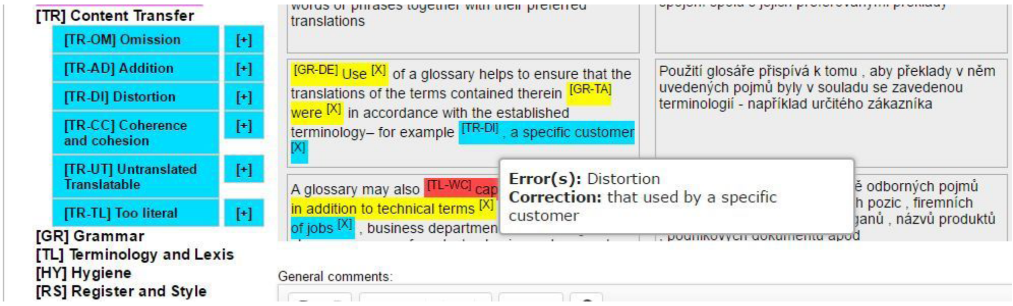
Before we get to a more detailed description of the individual statistics that can be generated in Hypal, it is important to describe the whole process of uploading student translations in Hypal, their further processing and the search in the compiled corpus. The process begins with setting up a *task.* The teachers choose a name and an identification code. Then they upload the source text (in the case of CELTraC a Czech text). More information about the source text can be added. The trainees are given the code, they receive a translation brief, and a deadline is set. Trainees can upload their translations themselves, mostly in MS Word or another word-processor format, directly into

Hypal by using their university student identity number and a password. All information about the student is thus stored automatically in Hypal in the form of metadata and can be used later for various searches and statistics (for more details, see Section 8). The student Hypal interface can be embedded in any webpage, including LMS Moodle (see https://hypal.eu/hypal) or accessed online [https://student.hypal.eu/].

Another option is to upload translations from previous years, retrieved from ELF, using students' identification numbers. Once the translations are in Hypal, the process of automatic paragraph alignment can begin. If necessary, it can be adjusted (see Figure 2). The next step then involves part-of-speech tagging and sentence alignment. These tasks are usually not very time-consuming since they are performed semi-automatically, separately for each translation. Then the corpus is ready to be error-annotated.

Figure 4. Icons for each step in corpus-compilation preparation: paragraph alignment, PoS tagging A,B, sentence alignment (M appears for Manual, A for Automatic), error tagging.

Figure 5. Manual error-tagging, *Content transfer,* correct solution is added.



translations

[GR-DE] Use [X] of a glossary helps to ensure that the translations of the terms contained therein [GR-TA] were [X] in accordance with the established terminology– for example [TR-DI] , a specific customer [X]

Použití glosáře přispívá k tomu , aby překlady v něm uvedených pojmů byly v souladu se zavedenou terminologií - například určitého zákazníka

A glossary may also [TL-WC] cap in addition to technical terms [X] of jobs [X] , business departmen

Error(s): Distortion
Correction: that used by a specific customer

ě odborných pojmů
h pozic , firemních
janů , názvů produktů

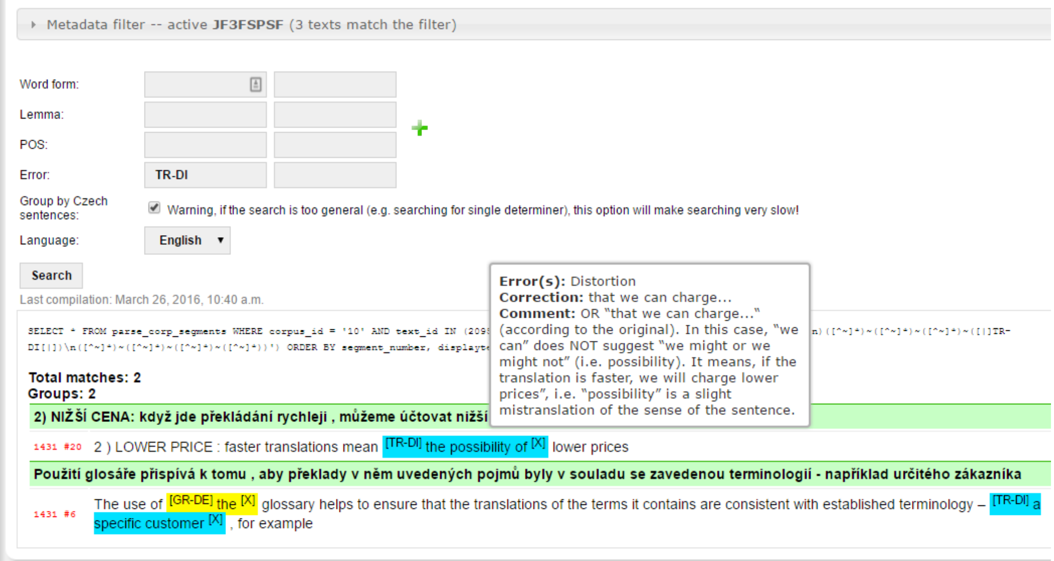
A corpus can be compiled with or without error-annotation. The corpus expands with each task added. It is also possible to add more translations to one task. Figure 5 shows errors marked in three colours – yellow for *Grammar*, turquoise for *Content Transfer* and red for *Terminology and Lexis*. All annotations are provided with correct versions and explanations, if necessary. These appear when hovering the mouse above the tag. Annotations can be removed by clicking on the X on the right.

8. Corpus Search

Once the corpus is compiled, various searches in the *Corpus Search interface* can be performed. The search interface, referred to as *Matrix search* is inspired by the CQL search syntax which is used for advanced queries in Sketch Engine. In Hypal, however, the user does not have to remember the syntax but forms the query by filling in the search table. Apart from error types, a particular word form, a lemma, a part of speech, or any sequence of those in either source or target language can also be searched for. Results can be grouped according to source language sentences.

As mentioned before, it is possible to use error abbreviations to search for error occurrences in either a particular translation, a course, or the whole corpus, using the settings in *Metadata filter*. If we search for a specific error, we get whole sentences, which means that sometimes other mistakes are included too. Figure 6 shows another example of TR-DI (distortion of content) with an explanation that, again, appears when hovering the mouse above the highlighted words.

Figure 6. Hypal corpus search interface – example of an error search.



Metadata filter -- active **JF3FSPSF** (3 texts match the filter)

Word form:

Lemma:

POS:

Error: TR-DI

Group by Czech sentences: ☑ Warning, if the search is too general (e.g. searching for single determiner), this option will make searching very slow!

Language: English

Last compilation: March 26, 2016, 10:40 a.m.

```
SELECT * FROM parse_corp_segments WHERE corpus_id = '10' AND text_id IN (209 n)([^~]+)~([^~]+)~([^~]+)~([|]TR-
DI[|])\n([^~]+)~([^~]+)~([^~]+)~([^~]+))') ORDER BY segment_number, displayt
```

Total matches: 2
Groups: 2

Error(s): Distortion
Correction: that we can charge…
Comment: OR "that we can charge…" (according to the original). In this case, "we can" does NOT suggest "we might or we might not" (i.e. possibility). It means, if the translation is faster, we will charge lower prices", i.e. "possibility" is a slight mistranslation of the sense of the sentence.

2) NIŽŠÍ CENA: když jde překládání rychleji , můžeme účtovat nižší

1431 #20 2) LOWER PRICE : faster translations mean [TR-DI] the possibility of [X] lower prices

Použití glosáře přispívá k tomu , aby překlady v něm uvedených pojmů byly v souladu se zavedenou terminologií - například určitého zákazníka

1431 #6 The use of [GR-DE] the [X] glossary helps to ensure that the translations of the terms it contains are consistent with established terminology – [TR-DI] a specific customer [X] , for example

It is also possible to search for a cluster of word forms, lemmas or parts of speech by adding another column to the search interface (Figure 6). If two expressions separated by a space in one window are entered, the programme will search for either of them. This particular feature can be used to look up translations of technical terms, various difficult passages in a translation, or just to see all occurrences of specific errors in various translations when discussing trainees' translations or pre-teaching a translation that has been used before with a different group of trainees.

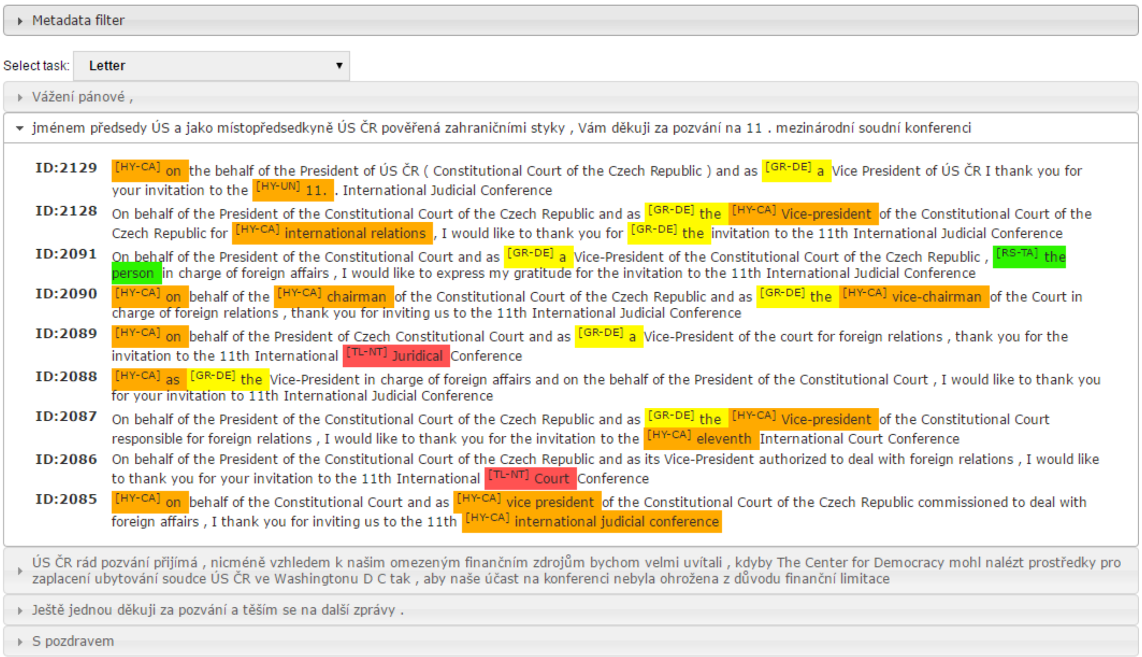
9. Teaching Tools

Teaching Tools are a recent addition to Hypal. They include two features, namely *Compare translations* and *Personal student reports*.

9.1. Compare Translations

As its own name suggests, this feature makes it possible to compare several translations of the same sentence. The *Metadata filter* allows the trainer to choose the text from the corpus and, if necessary, also the course in which the text was used, to be able to show the translations in the classroom, using e.g. a projector. In Figure 7, it is possible to see nine trainees' translations of the same sentence with error annotation. The abbreviation HY-CA stands for *capitalization*, GR-DE for *wrong determiner*, TL-NT refers to *term translated by non-term* and finally, RS-TA means *tautology, unnecessary repetition*.

Figure 7. Trainees' translations of the same sentence with error annotations.



▸ Metadata filter

Select task: Letter ▾

▸ Vážení pánové ,

▾ jménem předsedy ÚS a jako místopředsedkyně ÚS ČR pověřená zahraničními styky , Vám děkuji za pozvání na 11 . mezinárodní soudní konferenci

ID:2129 [HY-CA] on the behalf of the President of ÚS ČR (Constitutional Court of the Czech Republic) and as [GR-DE] a Vice President of ÚS ČR I thank you for your invitation to the [HY-UN] 11. . International Judicial Conference

ID:2128 On behalf of the President of the Constitutional Court of the Czech Republic and as [GR-DE] the [HY-CA] Vice-president of the Constitutional Court of the Czech Republic for [HY-CA] international relations , I would like to thank you for [GR-DE] the invitation to the 11th International Judicial Conference

ID:2091 On behalf of the President of the Constitutional Court and as [GR-DE] a Vice-President of the Constitutional Court of the Czech Republic , [RS-TA] the person in charge of foreign affairs , I would like to express my gratitude for the invitation to the 11th International Judicial Conference

ID:2090 [HY-CA] on behalf of the [HY-CA] chairman of the Constitutional Court of the Czech Republic and as [GR-DE] the [HY-CA] vice-chairman of the Court in charge of foreign relations , thank you for inviting us to the 11th International Judicial Conference

ID:2089 [HY-CA] on behalf of the President of Czech Constitutional Court and as [GR-DE] a Vice-President of the court for foreign relations , thank you for the invitation to the 11th International [TL-NT] Juridical Conference

ID:2088 [HY-CA] as [GR-DE] the Vice-President in charge of foreign affairs and on the behalf of the President of the Constitutional Court , I would like to thank you for your invitation to 11th International Judicial Conference

ID:2087 On behalf of the President of the Constitutional Court of the Czech Republic and as [GR-DE] the [HY-CA] Vice-president of the Constitutional Court responsible for foreign relations , I would like to thank you for the invitation to the [HY-CA] eleventh International Court Conference

ID:2086 On behalf of the President of the Constitutional Court of the Czech Republic and as its Vice-President authorized to deal with foreign relations , I would like to thank you for your invitation to the 11th International [TL-NT] Court Conference

ID:2085 [HY-CA] on behalf of the Constitutional Court and as [HY-CA] vice president of the Constitutional Court of the Czech Republic commissioned to deal with foreign affairs , I thank you for inviting us to the 11th [HY-CA] international judicial conference

▸ ÚS ČR rád pozvání přijímá , nicméně vzhledem k našim omezeným finančním zdrojům bychom velmi uvítali , kdyby The Center for Democracy mohl nalézt prostředky pro zaplacení ubytování soudce ÚS ČR ve Washingtonu D C tak , aby naše účast na konferenci nebyla ohrožena z důvodu finanční limitace

▸ Ještě jednou děkuji za pozvání a těším se na další zprávy .

▸ S pozdravem

This functionality can be used even prior to error-tagging. In this way it is possible to let trainees decide which translations they prefer and what problems they can spot. They may be asked to do the error-tagging themselves. Based on the most frequently occurring errors in a particular translation, it is possible to devise exercises to either remedy or pre-teach the problematic areas.

A more personalised feedback can be given to individual trainees by using the *Personal student reports* section of the *Teaching tools*.

9.2. Personal Student Reports

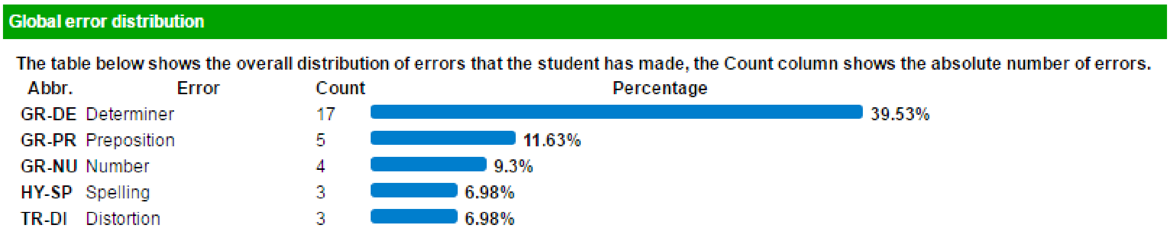
Personal student reports is one of Hypal's interfaces for the analysis of learner corpora. The interface helps the trainer perform global and longitudinal studies of errors made by a specific student. Non-anonymized personal student reports can be accessed by the teachers and the owner of the corpus. The owner of the corpus decides whether other corpus researchers working with the corpus can see the data as anonymized or not. Furthermore, students can view their own personal student reports.

9.2.1. Global error distribution

When opening this tab, a list of all students contributing to the corpus appears with the option to *Open statistics*. The number of texts included and annotated is listed for each student. The table in the first part of the *Personal student reports* section depicts the overall distribution of errors that a particular student has made, i.e. *Global error distribution* (see Figure 8).

Such statistics make it possible for translation teachers to spot the most common difficulties that a particular student encounters when translating specialized texts from Czech into English. When compared to global error statistics for the whole corpus (see Figure 9), the teacher can identify which translation-related difficulties are encountered by one particular student only and which, on the contrary, are faced by all trainees in general and thus need to be dealt with at a more global level. For example, it becomes clear by comparing figures 8 and 9 that determiners and punctuation seem to be global problems and need to be addressed in the curriculum. *Spelling* and *Number* (use of singular/plural), on the other hand, appear to pose a problem only to some students. These students might be provided with individual feedback to discuss the respective difficulties. The blue bars in Figure 9 show the statistics before filtering out *Positive feedback*, the red bars show the statistics after filtering out 835 *Positive feedback* annotation tags in CELTraC.

Figure 8. Five most frequent error types for one student.

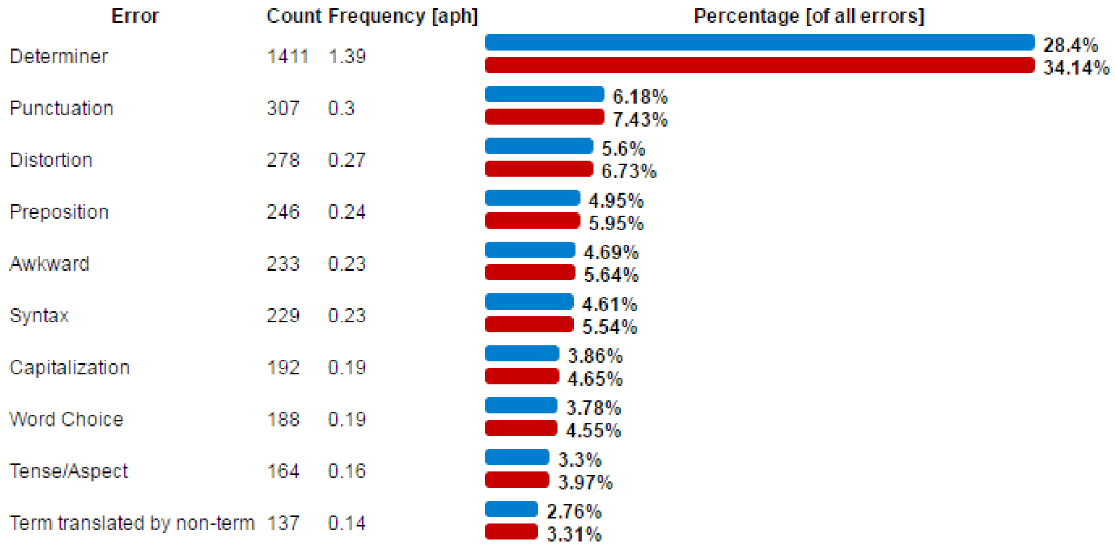


Global error distribution

The table below shows the overall distribution of errors that the student has made, the Count column shows the absolute number of errors.

Abbr.	Error	Count	Percentage
GR-DE	Determiner	17	39.53%
GR-PR	Preposition	5	11.63%
GR-NU	Number	4	9.3%
HY-SP	Spelling	3	6.98%
TR-DI	Distortion	3	6.98%

Figure 9. Ten most frequent error types in CELTraC with Positive feedback filtered out.



Error	Count	Frequency [aph]	Percentage [of all errors]
Determiner	1411	1.39	28.4% 34.14%
Punctuation	307	0.3	6.18% 7.43%
Distortion	278	0.27	5.6% 6.73%
Preposition	246	0.24	4.95% 5.95%
Awkward	233	0.23	4.69% 5.64%
Syntax	229	0.23	4.61% 5.54%
Capitalization	192	0.19	3.86% 4.65%
Word Choice	188	0.19	3.78% 4.55%
Tense/Aspect	164	0.16	3.3% 3.97%
Term translated by non-term	137	0.14	2.76% 3.31%

9.2.2. Error frequency divided into tasks

Apart from *Global error distribution* statistics, the *Personal student reports* section also offers charts showing a particular trainee's performance in individual tasks, i.e. *Error frequency divided into tasks* (see Figure 10).

Figure 10. Errors that the selected student made in individual tasks.

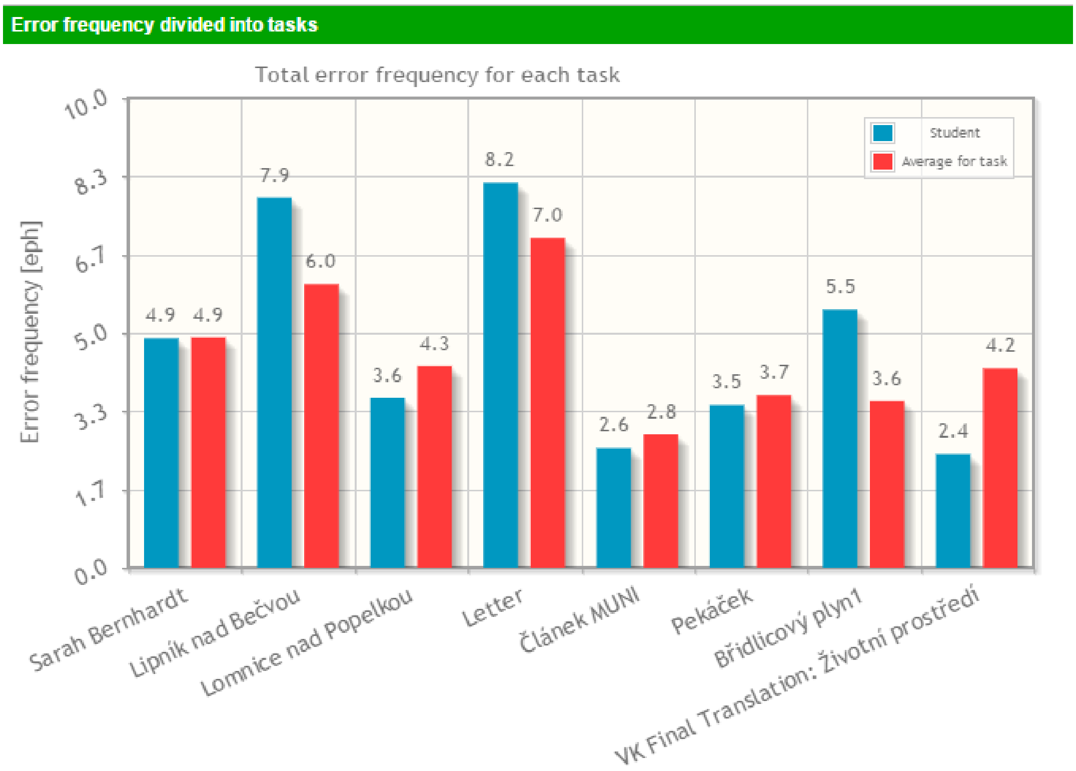
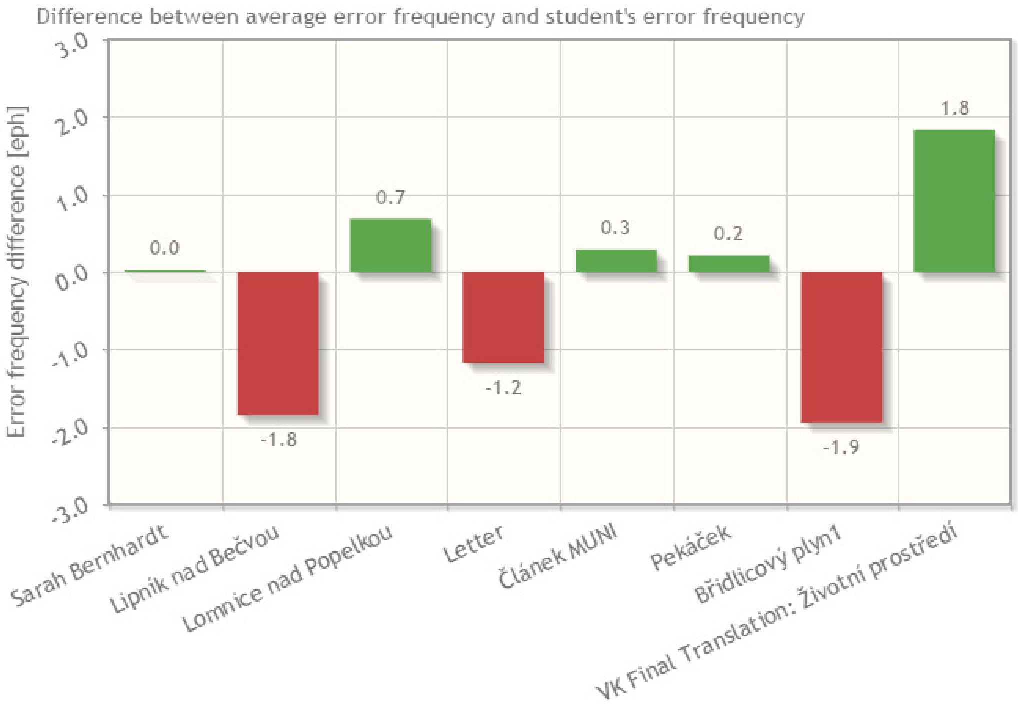


Figure 11. Difference between average error frequency and student's error frequency.



The blue bars in the chart above indicate that the student encountered most difficulties during the translation of the *Letter* task, making more than eight errors per one hundred words. This may be the result of the complexity and formality of the source text, suggesting that the student needs to extend his knowledge of the aspects of this particular genre. The chart further depicts that the student performs better during test translations – the best target text that the trainee produced was the final translation done in class. The relatively low number of errors (less than three per one hundred words) in the task may be, therefore, attributed to the trainee's motivation and the conditions in which the translation was produced.

Teachers are thus able to identify which types of source-text a particular student finds easy to translate and which, on the other hand, cause difficulties and therefore require special attention. In order to find out how a particular student performed in comparison with other trainees, the blue bars in the chart need to be compared with the red ones showing the average error frequency for individual tasks. The difference between the two bars reveals whether the student's performance in a particular task was better or worse than the average. The same data, but presented in a more explicit way, are shown in Figure 11.

It is obvious from the chart above that the trainee's translation behaviour is rather unstable – four out of eight translations are more or less average, three are considerably below-average, and the final task is significantly better-than-average. As indicated by the red bars, the worst translations that the student produced are *Letter, Lipník nad Bečvou* and *Břidlicový plyn 1.* Even though the *Letter* task contains the highest number of errors (see the blue bar for the task in Figure 10), the red bar in Figure 11 proves that the student actually performed worse during the translation of the *Břidlicový plyn 1* source text, for he/she made two more errors per a hundred words than most of the other trainees. These findings give clear evidence that trainees' performance varies according to source-text types.

9.2.3. Evolution of the error frequency over time

The charts in the third part of the *Personal student reports* section show the *Evolution of the error frequency over time* for a particular student, thus making it possible to monitor the development of that student's performance over time (see Figure 12).

The downward trend in Figure 12 clearly indicates that the translation performance of the selected student significantly improved over the course of two semesters. While in the translations submitted in autumn 2014 the student made more than five errors per a hundred words, in the translations produced six months later, in May 2015, the number of errors decreased to one per one hundred. Even though the downward trend is not perfectly steady, with a slight rise between March and April 2015, Figure 13 provides clear evidence that the student really made quite considerable progress.

The red line in the chart above represents the deviation from the average error frequency for the selected student. Whereas the student's first translation from November 2014 contains more errors than most translations of the same source text produced by other translation trainees, the three following translations that the student produced during 2015 are above-average, their quality gradually growing. Therefore, although the slight rise between March and April 2015 in Figure 12 suggests that the student's performance deteriorated over a month (the error frequency for the translation submitted in April is higher than for the one produced in March), the chart in Figure 13 disproves this, as the red line clearly shows that the student's translation produced in April is better-than-average.

Figure 12. Evolution of the overall error frequency for one student over time.

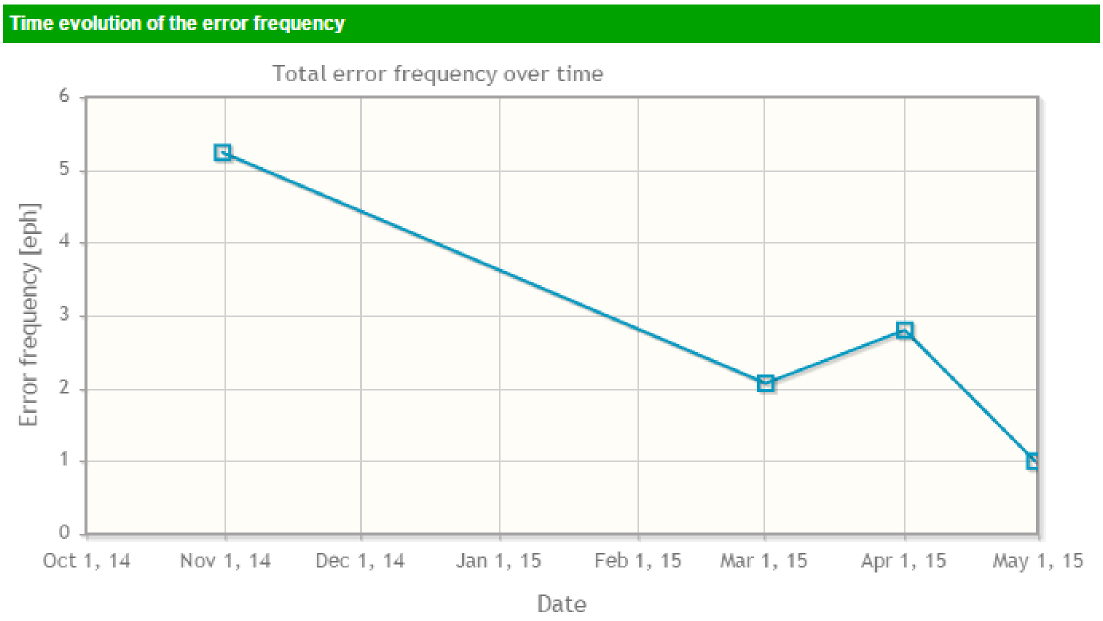
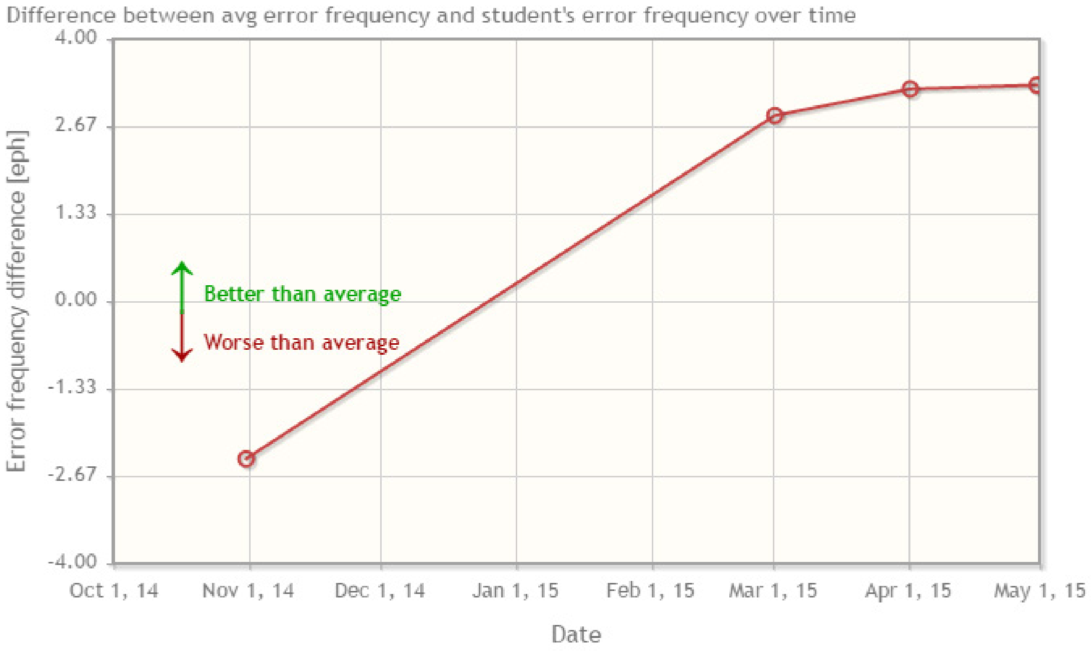
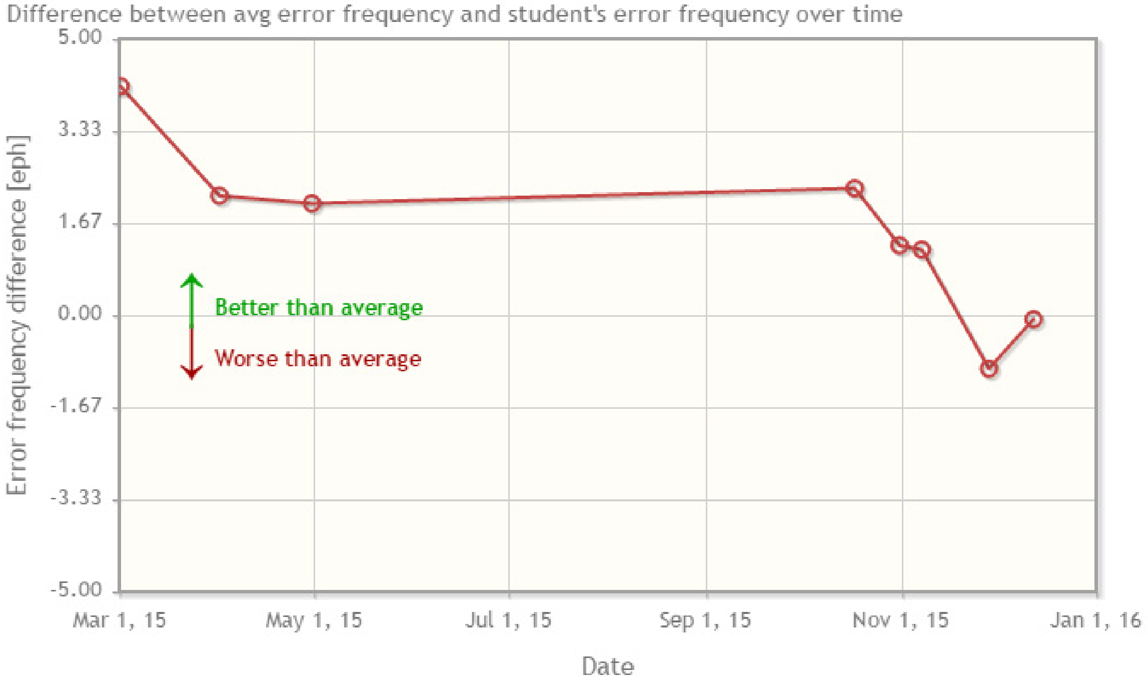


Figure 13. Time evolution of the deviation from average error frequency for one student.



Another example of the trainee's progress, this time, however, a negative one, is demonstrated by the line chart in Figure 14. Even though the error frequency for six out of eight translations that the student produced is lower-than-average, suggesting that the student possesses better translation skills than most of his fellows, the red line in Figure 14 clearly indicates that the trainee's translation performance was gradually deteriorating rather than improving.

Figure 14. Time evolution of the deviation from average error frequency for another student.

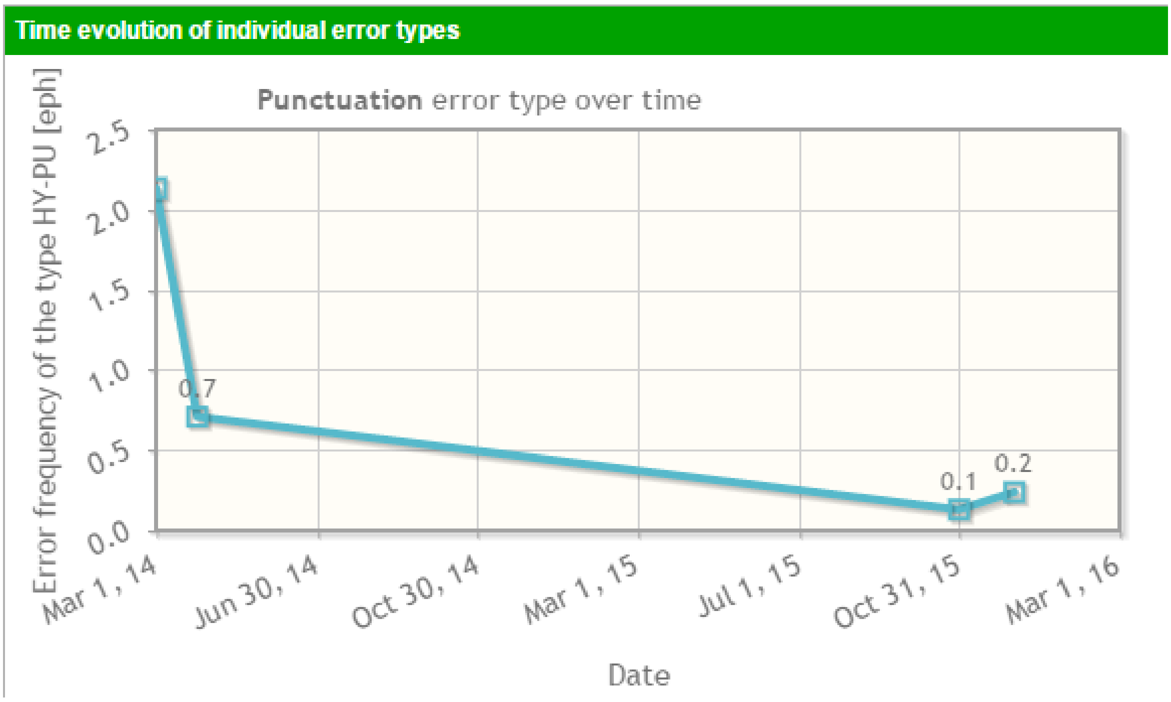


Since these conclusions are based on the analysis of eight translations produced over the course of one year, they are statistically more significant than those for the previous student. Nevertheless, before drawing any conclusions from the chart above it is necessary to take into consideration other factors and variables that might have influenced the findings, such as the type of the source text, inter-rater reliability, i.e. the degree of agreement among error-annotators, or the translation situation, i.e. whether the translation was done in class or submitted as a home assignment.

9.2.4. Time evolution of individual error types

The last part of the *Personal student reports* section offers charts depicting *Time evolution of individual error types* for a particular student. Studying these charts, a teacher may learn about the areas of translation in which trainees seem to have improved and those which are still causing difficulties. For instance, the line chart below illustrates a student's progress in the use of punctuation over the course of two years (see Figure 15).

Figure 15. Evolution of the frequency of Punctuation error type for one student.



While in March 2014 the selected student appears to struggle with punctuation, especially the use of commas, at the end of 2015 a relatively good command of this area of language is shown with less than one error of this type per one hundred words. Although the chart above is not an ideal example – the total number of punctuation-related errors that the student made is less than ten – it gives a clear idea of how the research into the evolution of individual error types might be beneficial to both translation trainers and trainees. Using the data from the charts, the trainer may identify a trainee as being weaker in a particular area of translation, notify them of the problem and provide them with individual feedback. In the *Corpus search* interface or in *Compare translations* as described above, the teacher can then retrieve specific examples of the trainee's use of punctuation in context, which helps them to see whether the trainee is indeed learning to use punctuation correctly. The empirical evidence illustrating the development of a trainees's learning can be used to encourage others to further develop their skills.

10. Research Tools

While the *Personal Student Reports* and *Compare Translations* interfaces are more likely to be used by teachers, there are also more advanced corpus analysis interfaces which were designed for more complex research. These are, in particular, the *Global Annotation Statistics* with versatile filtering possibilities and the *Corpus Search* interface. It is also possible to export the corpus from Hypal and search or analyse it using other software tools.

10.1. Global Annotation Statistics

The default view in the *Global Annotation Statistics* interface shows the overall distribution of all error types in the corpus (see Figure 9). However, the user can also filter the view based on the metadata, restricting the statistics to a specific course or students of a specific level (Bachelor or Master's). The researcher can also view which tokens are most frequently annotated by a specific error type.

10.2. Error-type Filter

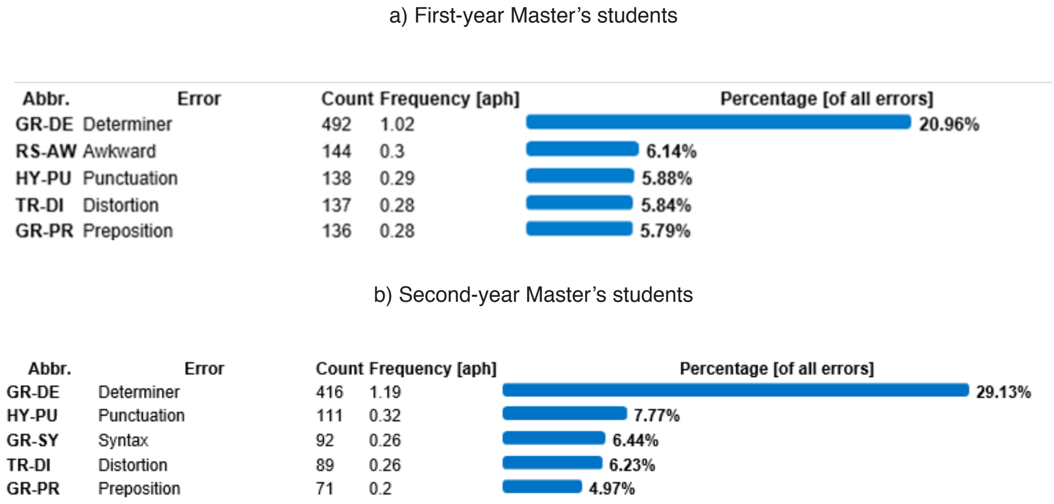
As described above, the error type filter makes it possible to restrict the view in *Global Error Statistics* to only specific error types or error categories. The *Error-type filter* can be combined with the *Metadata filter* described below, so that the corpus researcher can analyse how the frequency of selected error types depends on the metadata.

10.3. Metadata Filter

The *Metadata filter* is automatically generated for each corpus, based on the metadata which are available with each text. The corpus researcher defining the corpus can specify two classes of metadata, those that are filled in by teachers when creating an assignment (e.g. genre of the text, homework/in-class) and those that are automatically filled by the student when handing in the assignment (e.g. current year of studies, age, experience with professional translation…). Any combinations of the metadata can then be specified using the *Metadata filter* and only texts matching the criteria will be included in the annotation statistics. A simple example of the use of the *Metadata filter* is shown in Figure 16. The figure shows five most frequent errors made by students in the first and second years of Master's studies. It is apparent that the frequency of the *Awkward* error type decreased notably, as well as the error frequency of the *Preposition* error type, while the frequencies of all the other error types have remained approximately the same.

Taken together, these results suggest that some improvement can be achieved in the course of the training. It seems that the use of determiners (i.e. mostly articles) is a major problem for Czechs translating into English (see also Figure 9). The reason is that Czech has no articles, whereas in English articles are very frequent. Remedial exercises, pre-teaching, explaining and emphasizing the importance of correct usage can help improve the results.

Figure 16. Comparing five most frequent error types of first year (a) and second year (b) Master's students extracted from the CELTraC corpus.



a) First-year Master's students

Abbr.	Error	Count	Frequency [aph]	Percentage [of all errors]
GR-DE	Determiner	492	1.02	20.96%
RS-AW	Awkward	144	0.3	6.14%
HY-PU	Punctuation	138	0.29	5.88%
TR-DI	Distortion	137	0.28	5.84%
GR-PR	Preposition	136	0.28	5.79%

b) Second-year Master's students

Abbr.	Error	Count	Frequency [aph]	Percentage [of all errors]
GR-DE	Determiner	416	1.19	29.13%
HY-PU	Punctuation	111	0.32	7.77%
GR-SY	Syntax	92	0.26	6.44%
TR-DI	Distortion	89	0.26	6.23%
GR-PR	Preposition	71	0.2	4.97%

11. Conclusion

The aim of this paper was to introduce Hypal, a tailor-made software, designed to build parallel annotated corpora, which was developed as part of the teaching methodology within a Master's training programme. Students' participation in the pilot run of the Hypal project is significant and can be seen from the results of the research presented in this paper, as well as the numerous diploma theses written recently and listed below. In future, the newly updated error typology should be applied on the English into Czech translation learner corpus. It is necessary to transfer former trainees' translations that are stored in the e-learning system to Hypal. There are currently several student translations in Hypal that need error-tagging. Both teachers and new students need to be trained and continue with this work to prepare more data in order to improve the quality of teaching and the translations produced by trainees. It is expected that the English into Czech TLC will provide the trainers with more material that can contribute to better results in both teaching and research. The comparison of the two TLCs (CELTRac and the English-to-Czech translation learner corpus) and their error annotations should then bring more insights into the perceived nature of the impact that directionality has on translation processes and practices.

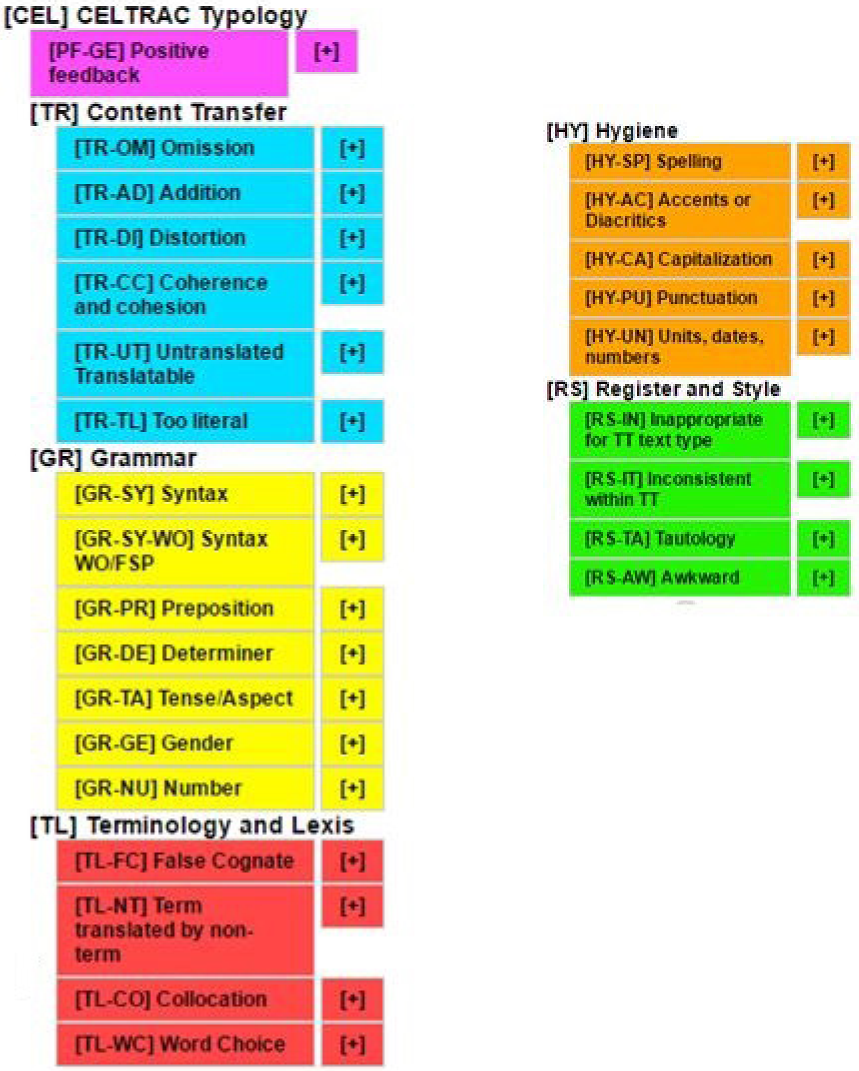
12. References

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

- Castagnoli, Sara, Ciobanu, Dragos, Kunz, Kerstin, Volanschi, Alexandra and Kübler, Natalie (2011). Designing a learner translator corpus for training purposes. In *Corpora, Language, Teaching, and Resources: From Theory to Practice.* N. Kübler (ed.). *221-248. Bern: Peter Lang.*
- Driscoll, Margaret (2002). Blended learning: Let's get beyond the hype. *E-learning* 1(4), 1-4.
- Espunya, Anna (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation* 48(1). 33-43. doi: 10.1007/s10579-013-9260-1.
- Fantinuoli, Claudio and Zanettin, Federico (2015). *New Directions in Corpus-Based Translation Studies*. Berlin: Language Science Press.
- Fictumová, Jarmila and Rambousek, Jiří (2014). Aus den Fehlern anderer lernen. Zur Entwicklung von annotierten Übersetzungslernerkorpora. In *Übersetzung als Kulturvermittlung. Translatorisches Handeln. Neue Strategien. Didaktische Innovation* (83-100). Bern: Peter Lang.
- Fictumová, Jarmila, Kamenická, Renata and Rambousek, Jiří (2014). Translation error typology for quality feedback: Czeching MeLLANGE through HYPAL. In *Proceedings of TIFO 2014: Translation and Interpreting Forum Olomouc*.
- Florén, Cecilia (2006). ENTRAD, an English Spanish parallel corpus created for the teaching of translation. In *Proceedings of TALC 2006: 7th Teaching and Language Corpora Conference*.
- Gale, William A. and Church, Kenneth. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1). 75–102.
- Gallego-Hernández, Daniel (2015). The use of corpora as translation resources: A study based on a survey of Spanish professional translators. *Perspectives* 23(3). 375-391.
- García Izquierdo, Isabel and Conde, Tomás (2012). Investigating specialized translators: Corpus and documentary sources. *Iberica* 23. 131-156.
- Jaaskelainen, Riita and Mauranen, Anna (2005). 5 Translators At Work: A Case Study Of Electronic Tools Used By Translators In Industry. In *Meaningful texts: the extraction of semantic information from monolingual and multilingual corpora*. Barnbrook, G., Danielsson, P. and Mahlberg, M. (eds.). London: Continuum
- Javora, Šimon (2016). *Defining an Error Typology*. Bachelor Thesis. Masaryk University, Brno.
- Kübler, Natalie (2007). The MeLLange learner translator corpus (LTC) [online]. <http://corpus.leeds.ac.uk/mellange/ltc.html>
- Kunilovskaya, Maria, Kovyazina, Maria and Ilyushchenya, Tatyana (2014). *Error-Tagging in Russian Learner Translator Corpus and Its Classroom Applications.* doi:10.13140/RG.2.1.2056.4640
- Obrusník, Adam (2013). *A hybrid approach to parallel text alignment*. Bachelor Thesis. Masaryk University, Brno. <https://is.muni.cz/th/356468/ff_b/?lang=en>

- Obrusník, Adam (2014). Hypal: A user-friendly tool for automatic parallel text alignment and error tagging. In *Proceedings of the 11th international conference Teaching and Language Corpora.*
- Štěpánková, Kristýna (2014). *Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus)*. Bachelor Thesis. Masaryk University, Brno.
- Sosnina, Elena P. (2006). Development and application of Russian Translation Learner Corpus. In *Proceedings of the 2006 Corpus Linguistics Conference.*
- Tamura, Akihiro, Watanabe, Taro and Sumita, Eiichiro (2014). Recurrent Neural Networks for Word Alignment Model. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. doi:10.3115/v1/p14-1138
- Tiayon, Charles (2004). Corpora in translation teaching and learning. *Language Matters* 35(1). 119-132. doi: 10.1080/10228190408566207.
- Uzar, Rafal and Walinski, Jacek (2007). Analysing the fluency of translators. Benjamins Current Topics, 135–145. doi:10.1075/bct.8.12uza.
- Varga, Daniel, Németh, Laszlo, Halácsy, Peter, Kornai, Andras, Trón, Viktor and Nagy, Viktor (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP 2005: Recent Advances in Natural Language Processing*.
- Votrubec (Raab), Jiří (2006). Morphological tagging based on averaged perceptron. In *Proceedings of WDS´06: Week of Doctoral Students*.
- Wurm, Aandrea (2013). Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE). *Trans-kom* 6(2). <http://www.trans-kom.eu/bd06nr02/trans-kom_06_02_06_Wurm_Eigennamen.20131212.pdf>

Appendix 1:
CELTraC Error Typology



[CEL] CELTRAC Typology
- [PF-GE] Positive feedback [+]

[TR] Content Transfer
- [TR-OM] Omission [+]
- [TR-AD] Addition [+]
- [TR-DI] Distortion [+]
- [TR-CC] Coherence and cohesion [+]
- [TR-UT] Untranslated Translatable [+]
- [TR-TL] Too literal [+]

[GR] Grammar
- [GR-SY] Syntax [+]
- [GR-SY-WO] Syntax WO/FSP [+]
- [GR-PR] Preposition [+]
- [GR-DE] Determiner [+]
- [GR-TA] Tense/Aspect [+]
- [GR-GE] Gender [+]
- [GR-NU] Number [+]

[TL] Terminology and Lexis
- [TL-FC] False Cognate [+]
- [TL-NT] Term translated by non-term [+]
- [TL-CO] Collocation [+]
- [TL-WC] Word Choice [+]

[HY] Hygiene
- [HY-SP] Spelling [+]
- [HY-AC] Accents or Diacritics [+]
- [HY-CA] Capitalization [+]
- [HY-PU] Punctuation [+]
- [HY-UN] Units, dates, numbers [+]

[RS] Register and Style
- [RS-IN] Inappropriate for TT text type [+]
- [RS-IT] Inconsistent within TT [+]
- [RS-TA] Tautology [+]
- [RS-AW] Awkward [+]

Definitions:

1. Content Transfer TR

Tag	Category	Definition	Example source	Example error in translation	Corrected translation
TR-OM	Omission	leaving out information from the translation	král Václav IV	King Wenceslas	King Wenceslas IV
			byla uctívána jako osmý div světa	was worshipped	was worshipped as an eighth wonder of the world
TR-AD	Addition	adding extra information to the translation	kabela	bread bag	bag
			vykopali jámu	dug a hole in the ground	dug a hole
			nejstarší synagoga na Moravě	the site of the oldest synagogue in Moravia	the oldest synagogue in Moravia
TR-DI	Distortion	altering the content of the original text in the translation	polovina 19. století	mid-20th century	mid-19th century
			židovská obec zanikla v roce 1942	the Jewish community disappeared in the year 1942	the Jewish community ceased to exist in the year 1942
TR-CC	Cohesion and Coherence	disrupting the flow of the translated text in any way	bohoslužby do roku 1941, pak využita ke skladovacím účelům	synagogue carried on with the services until 1941 after it was used as the storage building	the synagogue was used for worship until 1941, after which it became a storage area
TR-UT	Untranslated Translatable	leaving an expression untranslated that should be translated	rod Košíků z Lomnice	the family of Košíkové z Lomnice	the family of Košík of Lomnice
				the house of Košíkové of Lomnice	
TR-TL	Too literal	mechanically translating individual words instead of the whole expression	bohatý kostým	rich costume	opulent costume
			70. léta 19. století	seventies of the 19th century	the 1870s

2. Grammar GR

Tag	Category	Definition	Example source	Example error in translation	Corrected translation
GR-SY	Syntax	incorrect clause linking, redundant words	byla opravdu prezentována tak, jak si to představovala	she was truly presented as what she wished for	she was presented exactly as she wanted
GR-SY-WO	Syntax WO/FSP	incorrect word order errors	zákazníkům nic neúčtujeme	we do not charge anything to our customers	we do not charge our customers anything
GR-PR	Preposition	the use of incorrect, missing, or redundant prepositions	nejstarší synagoga na Moravě	the oldest synagogue on Moravia	the oldest synagogue in Moravia
			spadl do jámy	he fell down to the hole	he fell down the hole
GR-DE	Determiner	the use of incorrect, missing, or redundant articles or other determiners	nejstarší synagoga na Moravě	the oldest synagogue in the Moravia	the oldest synagogue in Moravia
			ten se vyřítil v plné zbroji, se sekerou	he, in his full armour, stormed out waving an axe	he, in full armour, stormed out waving an axe
GR-TA	Tense/ Aspect	incorrect use of verb tenses	Rod Košíků z Lomnice měl v erbu tatarského zvěda	The family of Košíkové of Lomnice have in their coat of arms the Tatar spy	The family of Košík of Lomnice had a Tartar spy in their coat of arms
			Synagoga prošla v dalších stoletích složitým vývojem	The synagogue has undergone a complex development in the following centuries	The synagogue underwent a complex development in the following centuries
			Loupežník okrádal pocestné	The highwayman was robbing wayfarers	The highwayman robbed wayfarers
GR-GE	Gender	assigning an incorrect grammatical gender (e.g. feminine instead of masculine)	byla ve skutečnosti malá, drobná, kostnatá žena s plochou hrudí	It was a short, petite, bony, flat-chested woman	She was a short, petite, bony, flat-chested woman
			Rod Košíků z Lomnice měl v erbu tatarského zvěda	In its coat-of-arms, the Košík House of Lomnice boasted a Tatarian spy	In their coat-of-arms, the Košík House of Lomnice boasted a Tatarian spy
GR-NU	Number	using singular instead of plural and vice versa	jedna z největších osobností Paříže	one of the most famous celebrity of Paris	one of the most famous celebrities of Paris
			Lomnice nad Popelkou a okolí	Lomnice upon Popelka and its Surrounding	Lomnice nad Popelkou and its Surroundings

3. Terminology and Lexis TL

Tag	Category	Definition	Example source	Example error in translation	Corrected translation
TL-FC	False Cognate	use of "false friends" – words that appear to be the TL equivalents of SL words but are not	knihy o secesi	books about the secession	books about Art Nouveau
			démonické fluidum	demonic fluid	demonic aura
			antická tragédie	antique tragedy	ancient tragedy
TL-NT	Term translated by Non-Term	using a non-standard translation of a term that has an established translation	obchodní cesta	commercial road	trade route
			lomničtí pánové	gentlemen of Lomnice	Lords of Lomnice
			sedlová střecha	V-roof	saddle roof
TL-CO	Collocation	using a non-standard expression instead of an established one	na přelomu let 1894-1895	on the verge of 1894 and 1895	at the turn of 1894 and 1895
			Spíše než scénu ze hry vidíme a cítíme celkový dojem	But rather than a picture from a play, we see and feel the overall impression	But rather than a scene from a play, we see and feel the overall impression
TL-WC	Word Choice	all terminology and lexical errors related to the incorrect choice of word not covered by the above categories	Vznik prvního plakátu Gismonda je obestřen muchovskými mýty	The creation of the first poster, Gismonda, is clouded in Muchian myths	The creation of the first poster, Gismonda, is shrouded in Muchian myths
			o tatarských zvědech	about Tatarian scouts	about Tartar scouts
			Maximum počtu židovských obyvatel	The maximum amount of the Jewish people	The maximum number of the Jewish people

4. Hygiene HY

Tag	Category	Definition	Example source	Example error in translation	Corrected translation
HY-SP	Spelling	misspelled words or typos	v Dalimilově kronice	in the Chornicle of Dalimil	in the Chronicle of Dalimil
			Loupežník okrádal pocestné	The robber stole form wayfarers	The robber stole from wayfarers
			Maximum počtu židovských obyvatel	The peek of the Jewish population	The peak of the Jewish population
HY-AC	Accents or Diacritics	omitting or adding diacritic marks	Lipník nad Bečvou	Lipnik nad Becvou	Lipník nad Bečvou
HY-CA	Capitalization	use of lowercase letters instead of uppercase and vice-versa	král Václav IV	king Wenceslas IV	King Wenceslas IV
			byzantskou ikonu	byzantine icon	Byzantine icon
			židovská obec	the Jewish Community	the Jewish community
HY-PU	Punctuation	the incorrect use of full stops, commas, colons, semicolons, quotation marks, dashes, hyphens, parentheses, and apostrophes	Mucha z ní vytvořil jakoby byzantskou ikonu	Mucha transformed her as it were into a Byzantine icon	Mucha transformed her, as it were, into a Byzantine icon
			v jihozápadním rohu	in the southwestern corner	in the south-western corner
			vládl král Václav IV	King Wenceslas's IV reign	King Wenceslas IV's reign
HY-UN	Units, Dates, Numbers	incorrectly formatted dates, units, or any other numbers in the target language	v průběhu 2. poloviny 15. století	in the course of the 2nd half of the 15th century	in the course of the second half of the 15th century
			v 70. letech 19. století	in the 1870th	in the 1870s

5. Register and Style RS

Tag	Category	Definition	Example source	Example error in translation	Corrected translation
RS-IN	Inappropriate for TT Text Type	using expressions not fitting the tone of the text type, such as using contractions in a formal text	Jisté je	It is sure that	It is certain that
			A nesejde na tom, vystupuje-li herečka v hlavní roli	And it doesn't really matter whether the actress stars	And it does not really matter whether the actress stars
RS-IT	Inconsistent within TT	minor tone violations with respect to text type	Mucha z ní vytvořil jakoby byzantskou ikonu	Mucha created somewhat Byzantine icon from her	Mucha created kind of a Byzantine icon from her
			v jihozápadním rohu historického jádra někdy ve 2. nebo 3. desetiletí 16. století	in the southwest corner of the historical center somewhere around 1520s or 1530s	in the southwest corner of the historical center around 1520s or 1530s
RS-TA	Tautology	unnecessary repetition	byla opravdu prezentována tak, jak si to představovala	she was pictured exactly just as she imagined	she was pictured exactly as she imagined
			na místě starší, patrně dřevěné předchůdkyně	taking the place of its, most likely wooden, predecessor from older times	taking the place of its, most likely wooden, predecessor
			byla rozšířena o severní dvoupodlažní trakt	it was expanded by adding the two-storey North Wing	it was expanded by the two-storey North Wing
RS-AW	Awkward	producing unnatural language	Naprosto přesně naplnil představy herečky	The poster fulfilled exactly the wishes of the actress	The poster exactly fulfilled the wishes of the actress
			Poté udeřili na hrad, aby loupežníka vyhnali	Afterwards they struck the castle to expel the robber out of the castle	Afterwards they struck the castle to expel the robber

Appendix 2:
List of Diploma Theses utilizing Hypal by Masaryk University students

(in chronological order)

- Obrusník, A. (2013). *A Hybrid Approach to Parallel Text Alignment*. Bachelor Thesis. Masaryk University, Brno.
- Šutariková, B. (2013). *Terminology of Annual Reports a Corpus Based Study*. Bachelor Thesis. Masaryk University, Brno.
- Štěpánková, K. (2014). *Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus)*. Bachelor Thesis. Masaryk University, Brno.
- Ambrožová, R. (2014). *Between True and False Friends: Corpus Analysis of Students' Translations*. Master's Thesis. Masaryk University, Brno.
- Pokorná, P. (2014). *An Analysis of Czenglish in an Error Tagged Learner Corpus.* Master's Thesis. Masaryk University, Brno.
- Knotková, M. (2015). *Parallel Texts as an Alternative Methodology in Investigation of Language of Space: The Case of English and Czech*. Master's Thesis. Masaryk University, Brno.
- Javora, Š. (2016). *Defining an Error Typology*. Bachelor Thesis. Masaryk University, Brno.
- Matusevich, I. (2016). *Quantitative Methods in Error Analysis on the Example of Errors in Academic Writing by Czech Advanced Learners of English.* Bachelor Thesis. Masaryk University, Brno.
- Dordová, L. (2016). *Translation into English: A Case Study of a Czech Museum.* Master's Thesis. Masaryk University, Brno.
- Heczková, E. (2016). *Czech-English Terminological Glossary for Caritas Czech Republic*. Master's Minor Thesis. Masaryk University, Brno.