Translation practice at the EU institutions: focus on a concordancing tool

Paola Valli
Università di Trieste

ABSTRACT

Translation has always played a major role within the European institutions because it provides the basis for democracy and communication among the Member States and between the EU and its citizens. The enlargements brought about changes in the internal organization of the institutions – including translation services and their workflow – to respond to the new challenge of accommodating 23 official languages. A greater need for translation support was met thanks to a growing number of shared tools and resources developed over time, such as centralised web-based applications and meta-search engines. This paper focuses on one specific tool available to translators working at the EU institutions, i.e. an internally developed multilingual concordancer. Concordancers are widely used by translators but little information is available about them in terms of tool evaluation or user behaviour. This article presents a PhD research project aimed to partly fill this gap by investigating the relationship between concordance searches (seen as manifestations of translation problems) and language combination within the EU translation services.

1. INTRODUCTION

This paper presents a PhD research project aimed at relating translation problems to language combination. The study draws on a dataset of hundreds of thousands of search logs retrieved from a specific web-based translation aid – a multilingual concordancer – which logs the queries submitted by EU staff translators.
In the first part of the article, background information about the translation services at the EU institutions (Section 2) and a description of the relevant Computer-Aided Translation (CAT) tools (Section 3) are provided. Section 4 gives a brief account of previous studies on translation tools and concordancers and sketches current scenarios where findings from this study may fit, while Section 5 deals in greater detail with the research project.

2. Translation services at the EU institutions

Given the political role multilingualism plays in maintaining a democratic basis within the European Union, dedicated translation services are available in a number of EU institutions. The largest is found at the European Commission (EC) that, with some 1700 translators, is known to be the largest translation service in the world. After the 2004 and 2007 enlargements, the European Parliament (EP) almost doubled in size, rising from 410 (Wagner et al. 2002: 15) to 750 translators (Kowalska 2010).

The main subject domains for translation at the EU institutions are politics, law and economics but just about any topic can be dealt with and some degree of specialization is often required in each service (Wagner et al. 2002: 44). For instance, the Treaties are translated by the translators at the Council of the EU; the EC and the EP deal mostly with legislative and legislation-related texts; the Economic and Social Committee and the Committee of the Regions are also involved in the legislative process because they present opinions on the draft legislation that have to be translated as well. Translators at the Court of Auditors are specialized in translating the annual and special reports about the EU financial resources and budget. Translators mostly deal with documents written in English, the most common source language even before the EU enlargement in 2004 (Dollerup 2001: 31).1 A special translation service can be found at the Court of Justice, where French is still the main working language. Staff translators have to be specialized in law and trained as lawyers because they translate judgements and orders and the opinions of the Advocates-General. A small number of lawyer linguists can also be found in other institutions, namely the Commission, the Parliament, the Council and the European Central Bank (Wagner et al. 2002: 15ff).

The European Commission deserves special attention as it has a complex internal organization and covers a wide range of text types. A whole Directorate-General is dedicated to translation (DGT) and a separate unit is specialized in Web publications. Documents can be divided into three main groups: incoming, outgoing and internal (DGT 2009a: 54,55; EC 2010: 47). Documents produced by DGT can be divided into two main types: legislation-

---

1 A special case is that of parliamentary questions to the Commission and/or the Council because these institutions can be addressed in the MEP’s native language, which makes the language range potentially much wider. According to the same principle, amendments to legislation can be presented in the MEP’s mother tongue. Both types of text are translated by the Parliament.
related, that is texts with specific EU terminology and usually available in all 23 official languages, and communication documents (printed material or website content) aimed at EU citizens, which will be adapted to the national context of each Member State (EC 2010: 47, 48). The enlargement brought about an urgent need for reorganizing the translation service of the EC, which went back to a language-based system integrating the pre-existing thematic organization (DGT 2009a: 47). This means that translators are grouped on the basis of their native language rather than by domain expertise.

The structure of the translation division at the European Parliament is more straightforward: Directorate A deals with support and technological services for translation, and Directorate B (“Translation and Terminology”) is divided into 23 language units, one for each official language, plus one Terminology Unit. The language regime at the EP is more balanced: there are no procedural languages – as is the case at the Commission with English, French and German – although a small group of languages are used for pivot translations, English undoubtedly being the most common. In 2009, about 80% of the source documents handled at DGT were written in English by non-native speakers (DGT 2009a: 56).

One very important aspect of translation for the European institutions regards the language policy in terms of directionality: translators are expected to translate into their mother tongue. This means that in each language unit, translators will be native speakers of the language they translate into.

3. Available CAT tools at the EU institutions

When it comes to translation support, the European Commission is arguably the most advanced and active among the EU institutions. The Commission has been using electronic resources for many decades. These resources are not intended exclusively for translators, and can also provide useful support (e.g. for documentation and administration purposes) to any person working at the EU institutions. Most of these tools have been developed internally or customized to meet the specific needs of the complex EU machinery (e.g. Euramis and Quest) and therefore are generally not available to freelance translators or outside contractors, to whom an increasing volume of translation work is outsourced. Back in 2004, Drugan (2004: 13) reported about multilingual document management and workflow at the EU institutions anticipating that 50% of all EU translations would be sent to external translators by 2007.

2 DGT translates the following document types: legal acts and preparatory documents; Commission decisions and communications; international agreements; policy statements; publications; technical studies; answers to written and oral parliamentary questions; correspondence; speeches and speaking notes; briefings and press releases; minutes; (financial) reports; working documents; internal administrative matters and staff information; scripts and captions for films and other promotional material; correspondence with ministries, firms, interest groups and individuals, and web pages (DGT 2009b: 3).
Technological aids that are more targeted to translation tasks can be attributed to one (or more) of the following categories (Lönnroth 2008): translation memories, terminology, machine translation, voice recognition software, workflow and transmission and web content management systems. Furthermore, libraries, the Internet and the intranet provide very useful resources for finding reference or background material and retrieving all necessary supporting documents.

Given that a considerable part of the written production of the institutions builds on existing legislation, documents to be translated generally do not contain just new or unseen text but they also quote previous legal texts, meaning that the translation task involves considerable searching, copying and pasting to ensure consistency. This is one of the reasons behind the large-scale adoption of computer-assisted translation software programmes, both commercial and proprietary.

Translation tools currently used at the Commission include terminology tools, translation memory technology (an EU-customised version of SDL Trados Translator’s Workbench), a customised rule-based Machine Translation (ECMT) engine and voice recognition systems, which are increasingly popular but generally limited to a small number of languages. Furthermore, Internet search engines have become an indispensable resource but many internal web-based tools have also the advantage of performing optimized searches in relevant domains, which allows a more efficient filtering of information and results.

Computer aids are made as user-friendly as possible, so that user interaction is reduced to a minimum. Training sessions for users are organised on a regular basis so as to improve tool adoption and acceptance, maximize software performance and bring users up to date with the latest developments. However, an excessively wide range of computer aids can also have negative effects for the translators and the workflow in general. Not everyone is tech savvy, and the impact of new learning curves can slow down the translation activity and increase the risk of inappropriate use of the tools, which can result in processing problems or lost information. These possible drawbacks notwithstanding, on-site interviews with DGT staff highlighted that “[…] the introduction of Euramis concordance features3 and IATE4 to build on and share terminology resources were generally mentioned as positive developments […]” (Drugan 2004: 19).

---

3 See sub-section 3.1.
4 IATE (Inter-Active Terminology for Europe) is the EU inter-institutional terminology database and contains entries in all the official languages. IATE has been used in the EU institutions and agencies since 2004 for the collection, dissemination and shared management of EU-specific terminology (http://iate.europa.eu) [last accessed: January 2011].
3.1 Euramis

EURAMIS stands for European Advanced Multilingual Information System. This system was originally developed by the European Commission in 1995 and is currently available to other institutions, namely the Council, the Court of Auditors, the Court of Justice, the Committee of the Regions, the European Economic and Social Committee, the European Parliament and the Translation Centre for the Bodies of the European Union. To access the system, users need to enter their log-in details so that correct access rights can be assigned, given that each institution has access to a set number of customized resources to protect sensitive content and make storage and retrieval more efficient.

Euramis consists of a series of centralized web-based applications for document search and retrieval, including a concordancing tool. A concordancer is commonly used by translators to query a repository of text segments to find the target language equivalent of a source text portion they entered as a search string. The advantage of using a concordancer over other translation tools is that text segments are displayed in their original context to help users make an informed choice when they look for the proper target language version. Euramis is referred to as a “central translation memory” (DG T 2009c: 10) because translators generally use a local Translation Memory (TM) with front ends such as SDL Trados Translator’s Workbench, Word or Trados TagEditor; the Euramis main repository is accessed in read-only mode. The Euramis concordancer is usually presented as a terminology search tool (Rusu 2009) with multilingual and multi-directional capability, so the underlying assumption is that terminological searches in context are the main reason for using a concordancer.

The user can submit a query either by opening the Euramis concordance page directly in the web browser and type or paste the string or by highlighting the relevant text portion and launch the search from a Word toolbar button. Searches can be performed using a simple or advanced concordance interface. The former only allows the selection of the source and target language and one or more database(s) while the latter has a number of additional filtering parameters.

Once the retrieval is completed, results are displayed in a new window. At the top, a box summarizes the chosen settings; below, a two-column table displays the results: source language on the left and the corresponding target version on the right. Only sentences (“segments”) containing the searched strings are shown and the text of the query is highlighted in bold red. If two or more segments are retrieved from the same document, they are grouped together under a common heading containing the metadata from the translation memory for document identification. Additional buttons are also available to open the whole document, download it locally or send feedback to the database manager, respectively.

A detailed description of the early Euramis algorithms for indexing and retrieval and the structure of the Translation Memory can be found in Blatt (1998a, 1998b). For details about the history of Euramis until the late 1990s, refer to Leick (1998).
3.2 Quest

Quest is a meta-search engine developed in the early 2000s by the European Commission to speed up the search process with simultaneous lookups across available databases and online resources. A new inter-institutional version of Quest was released in 2007.

Aside from Euramis, the other most prominent resources deal with terminology (IATE) and legislation (Eurlex) but other institution- and language-specific resources are also available and the user can select up to four different sources to be queried simultaneously. Just like the Euramis concordancer, Quest can be accessed online from the Quest webpage and the query can be typed directly in the search engine. Alternatively, the user can highlight a text portion in the text editor and launch Quest by using a toolbar button. The interface is basic: the user selects source and target language, a Search Profile and the resources to be included in the search.6

The result page comes in a two-column format. The left column lists all responsive databases on top. The first database to respond is automatically displayed as a nested web page in the larger right-hand column. By clicking on the relevant link, the user can change the displayed database. Euramis is very often the first database to be shown because of its fast response time. This means that, even though the translators launch their searches in Quest, they have in fact once again used Euramis. Overall, over half of the requests to the Euramis concordancer are submitted via Quest.

6 The number of available resources varies according to the institution but also the language pair.
4. Concordancing software

The Euramis concordancer and Quest are proprietary tools used in-house at the EU institutions, but today there are a number of concordancing tools publicly available that increasingly rely on collaborative content creation for the service they provide. There is evidence that corpus-based tools such as bilingual concordancers are popular among translators (Désilets et al. 2009) but there is hardly any information to be found about the way translators use them.

The structure of a bilingual (or multilingual) concordancer is relatively simple compared to other translation tools. It requires reference texts to be pre-processed so that the content can be quickly indexed and retrieved by the system. Reference documents usually come in the form of bi-texts or translation memories, which can be collectively (and loosely) labelled as aligned parallel corpora. All concordancing software works according to the same underlying principle: each search has to be manually launched by the user and then proposed solutions just need to be evaluated, accepted or ignored.

However, freely accessible online concordancers are not well suited for a systematic analysis of translators’ behaviour and problem solving strategies because there are too many uncontrolled variables to be dealt with simulta-

---

7 Examples are TAUS Data Association (http://www.tausdata.org/), MyMemory (http://mymemory.translated.net/), Glosbe (http://glosbe.com/) [last accessed: March 2011].

8 As the focus of this research is on translation aid tools supporting two or more languages, monolingual concordancers are not dealt with here.
Euramis and Quest, on the other hand, offer an ecologically valid and controlled environment for this exploratory study.

User behaviour in relation to the use of linguistic resources and tools has been investigated through ethnographic studies among Canadian professional translators carried out by the National Research Council of Canada (Désilets et al. 2009; Désilets, Brunette et al. 2008; Désilets, Farley et al. 2008). Special mention should be made to a tool called TransSearch,9 an English-French (-Spanish) concordancer, originally developed at the University of Montreal and currently commercialized by Terminotix, a Canadian software company, as it possibly represents the only object of academic research work carried out on a commercially available bilingual concordancer. Early work on TransSearch dates back to the early 1990s (Simard et al. 1993) and then continued in the 2000s with a focus on improving the concordancer with word-level alignment and a translation spotting functionality10 (Bourdaillet et al. 2010; Bourdailet et al. 2009; Huet et al. 2009a, 2009b). In the past decade, research on the tool mainly focused on the analysis of the concordancer search logs in order to shed light on the human translation process (Macklovitch et al. 2008; Simard & Macklovitch 2005; Macklovitch et al. 2000). In the first study on user behaviour (Macklovitch et al. 2000), researchers tried to elicit user information by means of a questionnaire. After having established that “users submit their queries in the natural course of their work, as they encounter translation difficulties” (Simard & Macklovitch 2005: 71), researchers focused on the linguistic nature of translation units seen as subconscious operational unit. A recent log analysis (Macklovitch et al. 2008) aimed at identifying the main types of translation problems that trigger the use of TransSearch.

Studies on TransSearch have been an important achievement given that a tool such as a stand-alone concordancer was usually only developed and used in academia for research or teaching purposes (Gavioli 1999, Scarpa 2006) and was not very well known to professional translators (Bowker & Barlow 2008: 8), even though common Translation Memory systems in fact integrate a concordancing functionality. The need for research on the concordancer as a translation tool has been recognised by authors studying how translators use CAT tools. In 2004, Lynne Bowker and Michael Barlow presented a comparative analysis of Translation Memory systems and bilingual concordancers (BCs) trying to fill the knowledge gap about the use of BCs by professional translators. A few years later, they noted a sustained lack of interest in BCs by other researchers: “[T]o the best of our knowledge, there have not been any detailed investigations that compare BCs to TMs” (Bowker & Barlow 2008: 2) – despite the existence of studies focused on the adoption on Translation Memory systems and related tools by translators (Lagoudaki 2006).

The knowledge gap in the field of concordancing tools has also been highlighted by the Canadian research group that advocated the need for greater

10 Also known as transspottinig. It can be defined as a feature that allows the system to highlight in the target text all corresponding translations to the searched string.
language coverage and for analyses of problems related to Language for General Purposes (Désilets et al. 2009) and pointed out that commercially available concordancers “have never been the object of scientific evaluation and publication” (Désilets, Farley et al. 2008).

5. The Research Project

The brief literature overview in the previous section shows how, in recent years, the research communities in translation process research, language resources and translation tools have grown interested in the way (professional) translators actually work. A number of significant findings have already been collected but there is also awareness of existing knowledge gaps, namely adequate language coverage and further investigation of translation problems that are not related to terminology.

The author’s PhD research project that will be presented in this section attempts to start filling these gaps by looking at a new set of data, which could shed light on the types of translation problems translators encounter on a daily basis. Furthermore, the data that will be used in the study may become a baseline or an additional form of user activity data for triangulating results in future experiments on translation process.

Euramis and the EU translation services at large provide an ideal test bed overcoming traditional practical problems that prevent a wide-scale cross-linguistic analysis, such as limited number of subjects, lack of comparable working conditions and translation tasks, not to mention variability in linguistic and translation resources. This study focuses on translation problems across multiple language combinations. Translation problems are assumed to be interpretable on the basis of the searches that translators performed via the Euramis concordancer.

Like many web-based tools, Euramis can log user activity. The search logs retrieved from the Euramis concordancer form the dataset for the analysis. A standard Euramis concordance log contains the following information: Date and Time stamp; Username; Institution Code (i.e. where searches come from); Source and Target Language(s); Searched Database; Execution Time; Maximum Number of Results (users can increase the default number); Number of Results; Sentence (searched text string); Search Interface used (Quest or Euramis) and Search Mode (simple or advanced). If the user selected the advanced search mode, the logs contain further details such as Search Method, DGs, Year(s), Year(s),

---

11 A major contribution to Translation Process Research comes from the findings on User Activity Data presented by the CRITT research group, which cannot be dealt with in this paper, due to space constraints. User Activity Data (UAD) are defined as “any kind of data which is consulted or generated by a translator during a translation session” (Carl 2009: 227).

12 On the other hand, the situation at the EU is quite special: Cosmai (2007: 77) makes an estimation of the total number or translators working internally of at least 3500 translators, while according to more recent figures (Kowalska 2010) the total is as high as 4500.

13 In the present study, this information was removed for privacy reasons.
Thanks to this information, log subsets can be created by filtering the queries submitted, for example, from one specific institution. Insights into the translation practices of each institution are important to adequately interpret the data, especially given that the search strings are provided out of the original context.

The Euamis concordancer receives an average of 32,000 queries per day (including weekends), which means that a very large volume of data can be collected in a reasonably short time span. An arbitrary time unit for the analysis has been selected (i.e. one full month\(^{15}\)), which amounts to 971,321 logged queries covering all 23 EU official languages. About 77% of the queries have English as the source language.\(^{16}\) This suggests that English as the source combined with the largest possible number of target languages (ideally all of them) is a good starting point for the analysis.\(^{17}\) In essence, the research project will study the Euamis concordance logs with a view to categorizing search strings and highlighting recurring search patterns across different language combinations with English as the source language.

Before embarking on a full cross-linguistic analysis, a pilot study will be undertaken to reduce the number of target languages, identify the best strategy for the analysis of the strings and automate data analysis as much as possible. For the pilot study, the English-Italian language pair has been selected, so that the test sample consists of almost 30,000 logs, roughly equivalent to the daily average of queries. Building on the findings of Désilets and his colleagues (2009) and the supposed use of Euamis concordancer for terminological purposes, a first macro-categorization of the strings will be attempted to assign the logs either to the Language for Special Purpose (LSP)\(^{18}\) or the Language for General Purpose (LGP) categories. According to the preliminary results, it seems possible to distinguish LSP strings from LGP strings in terms of their “substantiveness”. Thus, LSP strings typically occur in the form of noun phrases, whereas

---

\(^{14}\) As a means of comparison, TransSearch queries have been logged since 1997 and each log contains the following information: query submitted, number of hits produced, date and time, who submitted the query (i.e. source of the query), how results were displayed etc. (Macklovitch et al. 2008: 413; Macklovitch et al. 2000: 1204; Simard e Macklovitch 2005: 70). The similarities and/or differences between collected information in TransSearch and Euamis justify some of the methodological choices, because there are cases where the analysis carried out on TransSearch cannot be replicated in Euamis.

\(^{15}\) The chosen month is September 2010 because while the workload at the EC is fairly constant, the workload at the EP peaks during the Plenary sessions in Strasbourg and Brussels. In September, there were two long Plenary sessions and no holiday breaks. Therefore, translation activity is expected to be quite intense, especially compared to the summer months. Ideally, at least one additional month should be selected for comparison but the data volume is challenging enough, so this other analysis has been set aside for the time being.

\(^{16}\) According to official 2008 statistics, 72.5% of the source texts were drafted in English (DG T 2009b: 7) and, apart from French, all other languages were used very little as source languages.

\(^{17}\) Given the EU language policy outlined in section 2, we can assume that the selected target language in the searches is very likely to be the translator’s native language.

\(^{18}\) LSP is here broadly interpreted as a string that can be (semantically) ascribed to any EU-related domain.
LGP strings generally occur in the form of verb or prepositional phrases. Methodologies for a finer-grained clustering of the strings (e.g. into semantic domains) will be evaluated after completion of the first stage. They are expected to range from traditional computational linguistics approaches, such as POS tagging, to methods suitable for web logs, as described in the literature about Web Search Engine Log Analysis. The results of the study will help establish whether the language combination affects the type of translation problem encountered while highlighting search strategies and user behaviour patterns emerging from concordancer searches.
References


DGT (2009c) Translation Tools and Workflow, European Commission.


