How long is a piece of string? Concordance searches and user behavior investigated

Paola Valli
IUSLIT, University of Trieste
Trieste, Italy
paola.valli@phd.units.it

ABSTRACT

In the last few years, an ever-increasing number of translation resources have been moved into the cloud where shared repositories of language resources, such as Translation Memories, can be found. Access to these external TMs is generally granted through bilingual and multilingual concordancers. Concordancing tools enable users to be fully in control of the search process because searches are performed manually. This form of external support is used both by translators to solve translation problems and by non-translators to satisfy other information needs. This study analyzes concordance searches in terms of user behavior by drawing a comparison with Web search logs. It aims to identify the most common search strategies and types of interactions, recurrent search patterns, as well as possible issues that can negatively affect search efficiency. A large corpus of queries submitted by translators working at the European institutions will be analyzed across the EU official languages. Inferring user behavior directly from large volumes of authentic data can be used to gain further insights into translators' and general users' needs to improve currently available tools as well as develop new ones.

1 Why Concordancers (Still) Matter

Concordancers have been long known in corpus linguistics but they have also been available in Translation Memory Systems (TMS) since their early days to manually retrieve sub-segmental matches from a Translation Memory (TM). A concordancer is commonly used by translators to find a target language version of the source text portion they entered as a search string. The advantage of a concordancer over other forms of translation support is that matching text fragments are displayed in their original context to help users make an informed choice when they look for the paired target language version.

In recent years, translation resources have multiplied thanks to technological advances and joint efforts to share and integrate language resources. An ever-increasing part of the translation process has been moved into the cloud where collaborative platforms, shared repositories and increased automatization are possible. TMSs have been enhanced with machine translation and have automatized sub-segmental text re-use. At the same time, very large repositories of external Translation Memories have been stored in the cloud, giving translators and non-translators access to huge amounts of multilingual data. Such repositories can be accessed via a dedicated Web page and in some cases via a simple API or directly from the text editor. They can also come with a small desktop application, a Word macro or an add-on for integration into a TMS. External Translation Memories of this kind are usually queried online using a (multilingual) concordancer.

There are a few concordancing tools available online that offer different services. Some can be freely accessed (e.g. MyMemory [1], Glosbe [2], TAUS Data [3], Linguee [4]) and some require a paid subscription (TransSearch [5]), whereas others have been developed internally and can only be accessed via a (corporate) intranet (e.g. Euramis Concordancer at the EU). Such different levels of accessibility mean that there is a virtually wide range of user groups, from professional translators

[1] http://mymemory.translated.net/
[2] http://glosbe.com/
[3] http://www.tausdata.org/
[4] http://www.linguee.com/
[5] http://www.tsrali.com/

to the general public on the Web. In order to improve user experience for the different user groups, we need to know how people are currently using these tools. Unfortunately, there is hardly any data available that takes into account the new types of online linguistic resources and investigates user needs and translators' online search behavior.

Inferring user behavior directly from large volumes of authentic data – instead of relying just on the traditional data elicitation methods in Translation Studies – can be useful to identify translators' (un)expressed needs in real working conditions and consequently improve currently available tools or develop new ones. The purpose of the present study is to sketch the overall behavior of translators when using a concordancer to solve translation problems. Search patterns and recurring search strategies will be singled out focusing on a specific language pair when needed.

2 Concordance Searches as Information Needs

Concordance searches are generally launched manually, as opposed to the general trend for automatic translation proposals, where translators can only accept, edit or discard the suggested segment coming from a TM and/or a machine translation engine. This means that every concordance search is a deliberate choice that the translator made without any external prompting from the system.

A concordancer is seen as a form of support, accessed to solve some kind of translation problem. Previous research (Alves 1997) proposed a distinction between internal and external support to differentiate the adopted strategies by translators. External support comes into play when the translator resorts to any source of documentation to obtain information not immediately available to him/her (Alves & Liparini Campos 2009: 193) and the concordancer well fits this scenario. This study is not directly concerned with the selection of the type of support on the part of the translator. The underlying assumption is that the user, when accessing the concordancer, has already chosen this non-relational source as the one with the best trade-off between benefits and costs of attaining the desired information (Lu & Yuan 2011: 133). The nature of the problem may vary greatly from one search instance to the next and the same textual element might be associated with translation problems of varying magnitude, from double-checking a proposed solution (Buchweitz & Alves 2006: 258) to a reception problem (Krings 1986: 153).

Manually entering the text string, selecting filters and browsing through the results implies that the translator is fully in control of the search process. This approach to translation support tools involves specific cognitive activities such as developing a search strategy, balancing recall and precision or engaging with results assessment. Each concordance search should therefore be seen as a complex event where translation skills, search strategies, computer skills and translation-oriented Information Retrieval (IR) are combined.

As standalone concordancing tools are Web based, the model of a concordance search can be compared to that of a traditional Web query (Figure 1). An information need emerges from an ongoing task and the translator turns to a translation aid (the concordancer) for support. The expression *information need* will be used here to subsume all possible types of translation problems and can be defined as "the gap between people's current information and information sufficiency threshold" (Lu & Yuan 2011: 134). A query is then submitted to the system that retrieves matching results from a corpus of documents (in this case a large TM) and displays them to the user who can now choose to refine or change the search by submitting a new query to the system. Unlike classical IR, concordance searches generally *quote* portions of the source text, instead of containing search keywords put together by the user, as do traditional Web queries, because the ultimate goal of any concordance search is to find a target language version of a source text segment.

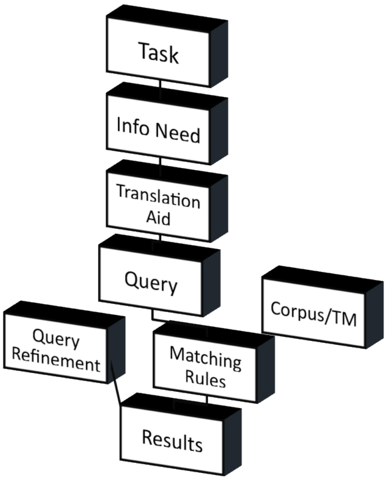

Figure 1 – Model for external support usage in translation (adapted from Broder 2002: 4)

If logs from freely available concordancers were to be systematically explored, a huge number of (uncontrollable) variables would have to be dealt with simultaneously. Just as concordance users may differ, so do concordancer interfaces. Concordancer interfaces can range from very simple to advanced search masks where a number of filters, including resource-specific ones, can be selected. Concordancers available to the general public tend to have a very simple interface with just the text box and the language selection mask and possibly some advanced options that can be activated if desired, whereas internally developed ones, such as the Euramis Concordancer, can be slightly more complex with both a simple and advanced interface. This affects user interaction with the tool, the way information is presented to the user and tool usability in general. A controlled but naturalistic environment was needed to perform an exploratory study on translators' search behavior that included a possibly homogeneous user group. A tool that could only be accessed via an intranet seemed a good starting point. Data collection for concordance searches can be carried out without interfering with the user thereby ensuring as much ecological validity as possible. Results from the analysis of a controlled environment may later be used to compare overall user behavior by including logs from freely accessible online concordancers.

3 Euramis and the Dataset

Euramis stands for European Advanced Multilingual Information System and was originally developed by the European Commission in 1995. It consists of a series of centralized Web-based applications for document search and retrieval, including alignment, machine translation and a concordancing tool. Euramis is the main repository of Translation Memories for several EU 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.

The Euramis Concordancer (henceforth simply "Euramis") is almost exclusively accessed by translating staff and covers all 23 official EU languages. By accessing Euramis via the EU intranet, translators can search shared Translation Memories as well as dedicated ones for their specific institution. During a search, the system logs data for each submitted query, including input string, tool settings and some additional metadata. The user can submit a query by opening the Euramis portal concordance page directly in the Web browser and type in or paste the string or s/he can highlight the relevant text portion and launch the search directly from the text editor. In this latter case, a metasearch engine (Quest) will be used that can simultaneously query up to four different resources.

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 the system was released in 2007. Quest has a very simple interface with minimal settings and is similar to a simple search in Euramis. Overall, well over half of the requests to the Euramis concordancer are submitted via Quest. As previously mentioned, Euramis also has an advanced interface where a variety of additional filters becomes available if the Web portal is used.

Once the retrieval is complete, results are displayed in a two-column table with paired source and target language segments. Only matching results are shown and the searched string is highlighted in the source column. If two or more segments are retrieved from the same document, they are grouped together under a common heading containing the document metadata, and the user can perform additional operations such as opening the document, downloading it or sending feedback about the found translation.

For this study, a month's worth of searches was collected from Euramis that covered virtually over 506 language combinations of all 23 EU official languages and amounted to some 970k queries. English was selected as the sole source language because it accounted for about 70% of the searches alone. Some cleanup and pre-processing operations were carried out to make the queries as consistent and noise-free as possible. Due to the exceedingly small size of the datasets for Gaelic and Maltese, these two languages had to be removed from the target language list. The final dataset amounted to 724,500 searches divided into 20 target language subsets whose distribution is summarized in Table 1.

TL	Queries Count	% (tot.)	TL	Queries Count	% (tot.)
NL	23,570	3.3%	SV	33,826	4.7%
PT	24,170	3.3%	RO	35,075	4.8%
ES	25,880	3.6%	CS	38,064	5.3%
FI	26,765	3.7%	HU	38,510	5.3%
EL	27,810	3.8%	SL	38,520	5.3%
IT	29,270	4.0%	PL	43,431	6.0%
LV	29,407	4.1%	LT	43,942	6.1%
SK	30,420	4.2%	ET	47,400	6.5%
DA	31,266	4.3%	DE	47,617	6.6%
BG	33,508	4.6%	FR	76,049	10.5%
Total	**724,500**	**100%**			

Table 1 – TL distribution of the final 724k dataset in ascending order.

4 Concordance Searches and Web Search Logs

The previously highlighted similarity between Web and concordance searching suggests that previous research on the interaction between users and Web search engines may provide useful insights and methodological approaches. Web log analysis is also known as Transaction Log Analysis (TLA), where a transaction log is defined as "an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine" (Jansen 2006: 408). TLA uses the "data collected in a transaction log to investigate particular research questions concerning interactions among Web users, the Web search engine, or the Web content during searching episodes" (Jansen 2006: 409).

In TLA, there are commonly three levels of analysis which will constitute the back bone of the present study: (i) *term level*, i.e. a string of characters delimited by a space or another separator, which in this case may be thought of as a "word" or, in corpus studies terms, a "token"; (ii) *query level*, i.e. a static analysis of the whole string of terms submitted to the engine, which will be later also referred to as "Problem Unit"; (iii) *session level*, where the dynamic evolution of the string in a limited time and multiple exchanges between a user and the system are examined.

4.1 Sessions

The three levels will be dealt with here in a top-down fashion. The first aspect to be analyzed is the search session. Based on the definition of *session* provided by Spink *et al.* (2009: 1361), i.e. "a series of queries submitted by a user and related interactions during an episode of interaction between the user and the Web search engine around a single topic," operational criteria were defined to comply with this definition, despite the fact the no explicit user information was available. Four criteria were identified, whose combination tries to ensure that the set of queries was indeed submitted by the same user, whereas the fourth also complies with the "single topic" feature, i.e. identity of information need. In order for a sequence of strings to be considered part of the same session, the following four requisites had to be met simultaneously:

1. The searches must come from the same institution
2. The searches must be submitted on the same day
3. The searches must be submitted within a two-minute time span[6]
4. The search string must have at least one word in common with the next string or the one after that (excluding a few stop words, likely to produce noisy results)

A customized PHP script was run on the subset for each language, splitting it into two files, one containing the search sessions, the other isolated queries (*spot searches*). Overall, about 36% of the concordance searches turned out to be part of a session but the percentage could be even higher if e.g. the search span of 3 strings were increased. The average session length calculated for each language ranged from 2.27 (FR) to 2.59 (BG) queries per search session. A comprehensive review of 25 years of research in the field of online IR systems (Markey 2007a) has similar results, i.e. that in the majority of studies the mean number of queries per end-user search session ranged between two and four queries per session (2007a: 1072).

Query reformulation (or query refinement) is the process by which a user modifies his/her previous search to either increase chances of obtaining results or fine tune the search and increase relevance in the output. Query reformulations are the building blocks of a search session and need to be looked at in order to study search strategies and information needs. Huang and Efthimiadis (2009: 77) report that about 28% of the daily 2 billion Internet searches are reformulations. References to existing taxonomies for query refinement can be found in Huang and Efthimiadis (2009: 78-79) who then move on to developing their own taxonomy of reformulation strategies. Some of these strategies are also applicable to concordance search sessions. A finer-grained classification was carried out here to better target the taxonomy to the special kind of data used, and a number of macro- and micro-reformulation strategies were identified (Table 2), each labeled with a code. The codes from five categories (A1, C1, C2, D1, D2) were automatically assigned to the relevant search sessions via a customized PHP script.

A	**RESUBMISSION**	D	**EXPANSION**
	A1. Repeated query		*D1. Left expansion*
	A2. Wildcards		*D2. Right expansion*
B	**FORMAL CHANGES**		*D3. Middle expansion*
	B1. Casing		*D4. Cross expansion*
	B2. Punctuation		*D5. Addition plural ‘s’*
	B3. Locale	E	**REPLACEMENT**
C	**REDUCTION**		*E1. Tense change*
	C1. Left trim		*E2. Paraphrase*
	C2. Right trim		*E3. Synonym/Antonym*
	C3. Middle trim		*E4. Word substitution*
	C4. Cross trim		*E5. Typo Fix*
	C5. Plural/Genitive ‘s’	F	**MIXED STRATEGY**

Table 2 – List of category codes employed to refer to a macro-category and each of its sub-types.

Results for Finnish show that categories C1 – "Left Trim" (25.62%) and C2 – "Right Trim" (23.05%) are by far the most used, followed by C3 – "Middle Trim" (4.23%) and C5 – "Removal of Plural 's'" (4.04%)[7]. Most remaining categories (including combinations of strategies for macro-category F) score lower than 4%. Overall, category C – "Reduction" is the most used strategy when searching with the concordancer. This finding seems to run counter to Web search behavior, where users were found to increase the length of the query in the course of the session, thus narrowing the information need (Huang *et al.* 2003 in Jansen *et al.* 2009: 1360). Translators, on the other hand, seem to prefer to start off by submitting a longer query and gradually trim away portions of

[6] A time span of two minutes was arbitrarily chosen after studying the logs because it was felt it was long enough to accommodate not only most of the actual sessions, which are usually terminated within less than a minute, but also sessions that took longer but were still almost consecutive. Longer time spans would increase the risk of noise in the results.

[7] Results for C1 and C2 are obtained with the script whereas percentages for C3 and C5 were obtained by a manual categorization of the sessions to test the PHP script. Percentages for C1 and C2 are in line with the overall mean for 20 languages of 23% and 21.4%, respectively.

it to increase recall over precision. This trend is confirmed by the contextual inquiry study carried out by Désilets *et al.* (2009) who reported that "[…] subjects seemed very adept at scanning a list of potential solutions, and rapidly sifting grain from chaff," particularly when the resource used was a corpus-based one, such as a list of Google hits or a bilingual concordancer. A search session therefore originates when the first search did not produce any useful results, either because retrieval failed (i.e. zero results) or because none of the (top) hits was satisfactory.

4.2 Queries

If search sessions can be said to illustrate the dynamic component in the interaction between the user and the retrieval system, a single query offers a static snapshot of a concordance search. The underlying problem-solving nature of the query implies a search strategy. The search strategy is the component of a concordance search that brings these search logs close to IR and Web queries and can be understood in terms of string length and tool setting selection whose combination directly affects recall and precision.

After calculating overall string length distribution across the whole dataset, it emerged that the submitted strings were all relatively short, ranging from one to five words, the vast majority containing only two words. A very similar distribution was found in a comparable bilingual concordancer, as shown in Table 3.

	TransSearch[8] **(6 years /7.2m)**	**Euramis (1 mnth/724k)**
Single-word queries	13.2%	13.83%
Two-word queries	39.6%	34.02%
Three-word queries	27.7%	20.33%
Four-word queries	13.0%	12.27%
Five-word queries	4.3%	6.66%
Six-word queries (+)	2.2%	12.90%
Total	**100%**	**100%**

Table 3 – Comparison of percentage distribution of query length between the TransSearch and the Euramis datasets.

These findings are also in line with results from Web studies, where mean query length was calculated to be about 2 to 3 terms per query (Silverstein *et al.* 1999: 8, Jansen & Spink 2000: 17, Johnson *et al.* 2006, Arampatzis & Kamps 2008, Jansen *et al.* 2009: 1365). Such homogeneous results may have several possible explanations but for the time being, reference will be made chiefly to the study by Azzopardi (2009: 560) concluding that "the communication with [the IR] system appears to be the most efficient […] when two to five query terms are used. […]." Overall, string length seems to suggest a desire for recall and efficiency in communicating with the system rather than precision.

In addition to changing the length of the query, the search can be fine-tuned by adjusting tool settings and selecting among available filters. Quest filters are limited to a few options: the resources to be queried and two search methods ('exact string' and 'all words'). On the other hand, Euramis comes with two interfaces, simple and advanced. In the simple interface, the user can only select a specific database (i.e. TM) and change the search mode from 'basic' to 'exact'. The advanced interface offers many more options to choose from: Search Method (with three options 'basic', 'exact' and 'any word'), Years, Requesting Service, Document Type, Document Number, Search Direction ('direct', 'reverse', 'indirect') and Maximum Number of Results to be displayed. It turns out that the only filter that is actively used is the 'Years' filter. In particular, users tend to select the most recent years (usually in pairs). Overall, advanced filters are selected once every 5 queries (20.6%), slightly more often in the case of sessions (25.22%) and slightly less in spot searches (17.8%). This may be partly due to the interaction with the metasearch engine that only works with the simple search mode. The intrinsic simplicity of the search process might also contribute to the overwhelming amount of basic searches, as found for Web logs (Markey 2007b: 1125). Advanced features of Web queries generally include a choice of Boolean operators but in fact "[e]nd users rarely use advanced system features and when they do, they are quite likely to use them incorrectly" (Markey 2007b: 1128). Once

[8] Macklovitch *et al.* (2008: 414)

again, concordance searches and Web queries resemble each other closely: concordance users submit short queries and hardly use advanced features.

If the search strategy can be seen as the problem-solving part of the search, the searched text portion constitutes the problematic element that caused the interruption of the translation task, i.e. the Problem Unit. The concept of Problem Unit has been used in Translation Studies in conjunction with process and pedagogy (e.g. Kilary 1995: 75) but in fact the notion of *problem* in Translation Studies has been used with three different senses (Toury 2002): (i) related to the concept of "translatability", (ii) linked to instances of translation in product-oriented studies and (iii) in process-oriented studies, where translation is observed as it unfolds. The present analysis focuses on the "nature" of a problem unit as understood in translation process-research, i.e. marked by interruptions. To this end, three levels of analysis were identified (i) size, (ii) content, (iii) linguistic form (Figure 2).

Size is a property that has already been dealt with in the discussion over search strategy (see Table 3), with queries turning out to consist mainly of multi-word units (MWUs).

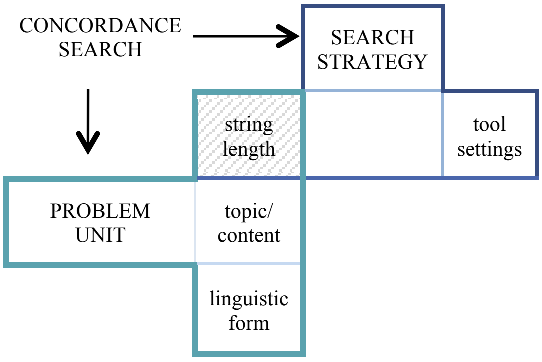

Figure 2 – *Breakdown of a concordance search into its two main components and related sub-components. "String length" is a shared feature*

The next step will be to try and categorize queries according to their content. In research on Web behavior, different clustering techniques have been presented, such as co-occurrence of terms (Ross & Wolfram 2000), manual classification in topical categories (Jensen & Booth 2010) and autocategorization based on feature terms (i.e. seed terms manually categorized into a predefined taxonomy) (Pu *et al.* 2002).

Unfortunately, none of the employed taxonomies could be successfully applied to the present study and manual classification and category creation were avoided. A readily available taxonomy covering all the EU-related activity fields was found and chosen as reference. This taxonomy (EuroVoc[9]) is officially known as a "multilingual thesaurus" because all labels are available in all official EU languages and more. It is organized into a two-tier hierarchical structure with numbered fields (Table 4) and microthesauri. Using a customized Perl script each string was compared with the available descriptors and whenever a match was found it was labeled with the relevant field code.

Eurovoc Fields	
POLITICS (04)	EMPLOYMENT & WORKING CONDITIONS (44)
INTERNATIONAL RELATIONS (08)	TRANSPORT (48)
EUROPEAN COMMUNITIES (10)	ENVIRONMENT (52)
LAW (12)	AGRICULTURE, FORESTRY & FISHERIES (56)
ECONOMICS (16)	AGRI-FOODSTUFFS (60)
TRADE (20)	PRODUCTION, TECHNOLOGY & RESEARCH (64)
FINANCE (24)	ENERGY (66)
SOCIAL QUESTIONS (28)	INDUSTRY (68)
EDUCATION & COMMUNICATIONS (32)	GEOGRAPHY (72)
SCIENCE (36)	INTERNATIONAL ORGANISATIONS (76)
BUSINESS & COMPETITION (40)	
[EUROJARGON (00)] – addition to EuroVoc	

Table 4 – Official Eurovoc fields in ascending order according to official field code

The purpose of this analysis is twofold. On the one hand, an attempt was made to automatically classify strings into the two categories of Language for General Purposes (LGP) problems and Language for Special Purposes (LSP) problems. The resulting distribution was compared to the one manually obtained in the study by Désilets *et al.* (2009), where LSP and LGP problems occurred in about the same proportions, the former accounting for about 41% of all consultations. Overall, almost 58% of the queries received at least

[9] http://eurovoc.europa.eu/drupal/

one EuroVoc label, which seems to be in line with the previous findings.

On the other hand, EuroVoc was useful to obtain both a finer-grained classification of LSP queries and the distribution of queries across all domains to identify the potentially most problematic queries. Results showed that the domains concerning European Communities (10; 19.8%), finance (24; 14.42%), politics (04; 14.3%) and law (12; 14.3%) were the most populated (Figure 3).

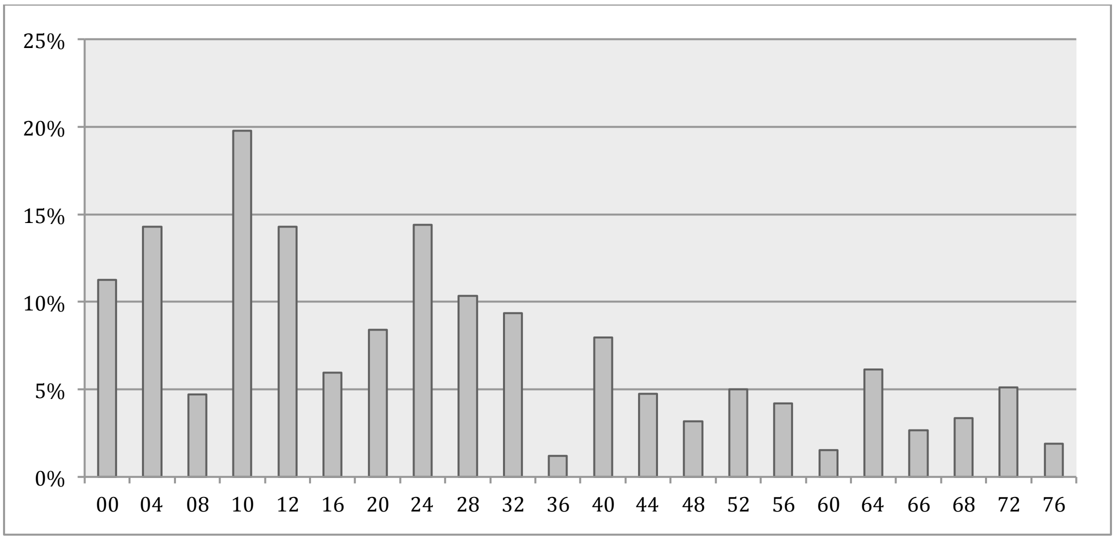

Figure 3 – Percentage distribution of queries with at least one EuroVoc match (ca. 58% of 724k) across all domains.

The analysis of the linguistic form is the most problematic due to the lack of operational categories that can be effectively converted into formal language and the extreme variability of the strings. These issues eventually restrict the analysis almost exclusively to the qualitative aspect. Due to space constraints, the analysis will be limited to a small qualitative sampling study at the lowest level of the Web hierarchy – the term.

4.3 Terms

Strictly speaking, only single words are considered at term level, i.e. the equivalent of tokens in corpus linguistics. For this quick overview of terms, the focus will be directed to single-word queries. A frequency list of all strings in the dataset shows that there are as many as 27 single-word queries in the 100 most frequent queries and about half (13) are acronyms. At first glance, most of the remaining terms can be labeled as abstract nouns ('stakeholders', 'governance', 'expertise', 'accountability', 'enforcement') while the rest falls into the tentative categories of technical (financial) terminology ('leverage', 'derivatives')[10], *-ing* forms ('mainstreaming', 'networking') and polysemous or ambiguous terms ('grant', 'benchmark').

In this last part, special attention will be devoted to acronyms that can also be found in their spelled-out form and, just like their full versions, can be considered a type of named entity. Named entity queries are a very popular type of query and studies report that at least 20-30% of queries submitted to some search engine were named-entity queries (Yin & Shah 2010: 1001). In terms of predicting the underlying information need of a translator, named entities (including acronyms) can possibly be considered one of the most transparent searches, i.e. the translator is after the corresponding name or acronym in the target language because there is usually a 1:1 correspondence between source and target language. This

[10] Without any further context, these terms could easily be attributed to a different domain than finance but given the results of the EuroVoc analysis chances are that they are indeed financial terms.

type of search can only be successful if the database already contains the named entity in question, as opposed to other types of searches, where (low) fuzzy matches may still help.

Some of the known issues in named entity recognition are ambiguous capitalization, semantic and structural ambiguity. For example, the named entity that corresponds to the official title of the Vice-President of the European Commission, Catherine Ashton, High Representative of the Union for Foreign Affairs and Security Policy, was found in 56 different versions in the dataset alone, ranging from 2 to 17 words per string. Also, the term 'erdf' had a frequency of 155 (rank 9) while its spelled out form ('european regional development fund') was found 115 times (rank 29). From a translator's point of view, the frequency of this named entity should in fact be the sum of the two, i.e. 270 (more than the top ranked string), because spelled-out forms and acronyms are sometimes interchangeable and at other times appear next to each other. Both named entities used in the examples occur rather frequently in EU texts, so finding results should be easy in this particular case. However, the fact that so many acronyms (and named entities) appear so high in the frequency list of translation problems suggests that there is room for more target help with this type of query.

5. Conclusion

This study has focused on concordance searches submitted by EU staff translators to an internal translation resource (the Euramis multilingual concordancer) in order to study user behavior and search trends.

From the very beginning, concordance searches closely resembled Web queries and most of the results confirmed the similarities. Concordance searches tend to be very short (2-4 words) and search sessions usually do not exceed 2-3 queries. Users do not seem particularly concerned about the precision of results and tend not to apply any filters to their searches.

When it comes to search sessions, the main trend is to progressively shorten the query instead of lengthening it as in Web searches. This is possibly linked to the ultimate information need of a translator, i.e. finding a target language version for the problematic source text portion, where recall is more "useful" than precision.

In terms of content, no direct comparison could be made between Web and Euramis searches due to the specificity of the EU context. A significant group of strings did not belong to any specific domain (LGP problems). On the other hand, LSP searches highlighted EU-related queries as the most popular domains for queries followed by finance and law.

Acronyms and, by extension, named entities turned out to be a much sought-after item for which the current systems lack specific support.

A more detailed study of all these aspects is being carried out in the author's PhD research project where the aim is to create baseline results with a view to extend the analysis to concordance searches submitted by the general public to freely available concordancers.

Acknowledgments

The author would like to thank the IT unit of the European Commission's Directorate-General for Translation (DGT) in Luxembourg and in particular the Euramis project team for providing the data and their invaluable support and continued assistance. A special thank goes to prof. Daniel Hardt (Copenhagen Business School) and Mr. Antonio Farina (Translated s.r.l.) for their help with the scripts.

References

Alves, F. (1997). A formação de tradutores a partir de uma abordagem cognitiva: reflexões de um projeto de ensino. *TradTerm*, *4*(2), 19–40.

Alves, F., & Liparini Campos, T. (2009). Translation technology in time: investigating the impact of translation memory systems and time pressure on types of internal and external support. In S. Göpferich, A. L. Jakobsen, & I. M. Mees (Eds.), *Behind the*

Mind. Methods, models and results in translation process research, Vol. 37. Copenhagen: Samfundslitteratur, 191-218.

Arampatzis, A., & Kamps, J. (2008). A Study of Query Length. *SIGIR'08. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY: ACM, 811-812.

Azzopardi, L. (2009). Query side evaluation. *SIGIR '09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. New York, NY: ACM, 556-563.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3–10. Retrieved from http://www.sigir.org/forum/F2002/broder.pdf

Buchweitz, A., & Alves, F. (2006). Cognitive Adaptation in Translation: an interface between language direction, time, and recursiveness in target text production. *Letras de Hoje*, *41*(2), 241–272.

Désilets, A., Melançon, C., Patenaude, G., & Brunette, L. (2009). How translators use tools and resources to resolve translation problems: an ethnographic study. *MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT*. Ottawa, Ontario, Canada. Retrieved from www.mt-archive.info/MTS-2009-Desilets-2.pdf

Huang, C.-K., Chien, L.-F., & Oyang, Y.-J. (2003). Relevant term suggestion in interactive web search based on contextual information in query session logs. *Journal of the American Society for Information Science and Technology*, *54*(7), 638–649.

Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. *Proceeding of the 18th ACM conference on Information and knowledge management CIKM 09*. New York, NY: ACM, 77-86.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, *28*, 407–432.

Jansen, B. J., & Booth, D. L. (2010). Classifying Web queries by topic and user intent. *CHI 2010: Work-in-Progress*. Atlanta, GA. Retrieved from http://faculty.ist.psu.edu/jjansen/academic/jansen_user_intent.pdf

Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during web searching. *Journal of the American Society for Information Science and Technology*, *60*(7), 1358–1371.

Jansen, B. J., & Spink, A. (2000). Methodological approach in discovering user search patterns through web log analysis. *Bulletin of the American Society for Information Science*, (October/November), 15–17.

Johnson, D., Malhotra, V., & Vamplew, P. (2006). More effective web search using bigrams and trigrams. *Webology*, *3*(4). Retrieved from http://www.webology.ir/2006/v3n4/a35.html

Kiraly, D. C. (1995). *Pathways to Translation. Pedagogy and Process*. Kent: Kent State University Press.

Krings, H.-P. (1986). *Was in den Köpfen von Übersetzern vorgeht*. Tübingen: G. Narr.

Lu, L., & Yuan, Y. C. (2011). Shall I Google it or ask the competent villain down the hall? The moderating role of information need in information source selection. *Journal of the American Society for Information Science and Technology*, *62*(1), 133–145. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.21449/pdf

Macklovitch, E., Lapalme, G., & Gotti, F. (2008). TransSearch: What are translators looking for? *AMTA 2008* - Hawaii, USA, 412-420. Retrieved from http://www.mt-archive.info/AMTA-2008-Macklovitch.pdf

Markey, K. (2007a). Twenty-five years of end-user searching. Part 2: Future research directions. *Journal of the American Society for Information Science and Technology*, *58*(8), 1123–1130.

Markey, K. (2007b). Twenty-five years of end-user searching. Part 1: Research findings. *Journal of the American Society for Information Science and Technology*, *58*(8), 1071–1081.

Pu, H.-T., Chuang, S.-L., & Yang, C. (2002). Subject categorization of query terms for exploring Web users' search interests. *Journal of the American Society for Information Science and Technology*, *53*(8), 617–630.

Ross, N. C., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science and Technology*, *51*(10), 949–958.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum Newsletter*, *33*(1), 6–12. Retrieved from http://www.sigir.org/forum/F99/Silverstein.pdf

Toury, G. (2002). What's the problem with "Translation Problem"? In B. Lewandowska-Tomaszczy & M. Thelen (Eds.), *Translation and Meaning Part 6*. Maastricht: Hogeschool Zuyd, Maastricht School of Translation and Interpretation, 57-71.

Yin, X., & Shah, S. (2010). Building a taxonomy of web search intents for name entities queries. *WWW '10 Proceedings of the 19th international conference on World Wide Web*. New York, NY: ACM, 1001-1010. Retrieved from http://research.microsoft.com/pubs/120889/fp0700-yin.pdf