A Platform for the Empirical Analysis of Translation Resources, Tools and their Use

David Day, Galen Williamson, Alex Yeh, Keith Crouch, Sam Bayer, Jennifer DeCamp, Angel Asencio, Seamus Clancy and Flo Reeder

The MITRE Corporation
202 Burlington Rd., Bedford, MA 01730, USA
{day,gwilliam,asy,kcrouch,sam,jdecamp,asencio,sclancy,freeder}@mitre.org

Abstract

We have developed a software framework that will support experiments to explore the role of translator resources and tools in the performance of translation and translation-related activities. This software environment brings together a wide range of resources and tools within a single work environment that has been instrumented to measure the actions of the translator. In this paper we present an overview of the system that has been developed and describe the kinds of experiments that we intend to conduct. The platform provides detailed logs for most of the actions taken by a translator using the tool suite. We intend to use the data collected from controlled experiments to explore a number of questions, such as how resources and tools effect the productivity and quality of translators depending upon their level of experience, the texts on which they are working, the time constraints imposed on their work, and the mix of resources/tools made available.

1. Introduction and Overview

There is an increasing focus on the potential for linguistic resources and computational tools to enhance the productivity and quality of human translation. Computational linguists and tool developers are rushing forward to create a wide variety of tools and resources that they argue will provide translators, especially those without the benefit of complete mastery of the craft or working in terminologically demanding domains, with labor saving and/or insightful ways of approaching the process of transforming texts across the linguistic gulf of two languages and cultures. At the same time some experienced translators are dubious about the value of some of these proposals, arguing that the translation task is dominated by the very human process of comprehending the author's intent and finding an adequate choice of words for capturing this intent in the target language.

As part of the larger effort of coming to a greater understanding of how translators actually perform their tasks, and more specifically to identify the extent to which emerging computational tools and resources can influence the human translation task, we have developed an experimental prototype for performing empirical analyses of the translation task and the use of ancillary materials and tools. This platform has incorporated a range of resources and tools within an instrumented environment, by which we hope to record a number of relevant user behaviors in the process of performing translation and so help in understanding the true role for some of these contributions.

The emphasis of this environment is on exploring the potential advantages of tightly integrating emerging automated computational aids, This effort is not at the present time attempting to address some of the more fundamental aspects of translation and the associated cognitive processes, which have been and are continuing to be explored by others (e.g., Tirkkonnen-Condit, 1986; Danks, et al, 1997). This focus has meant that some capabilities that would support such explorations have been left out. For example, the level of detail we currently capture within the activity logs (details are described later in the paper) are suitable for addressing the specific hypotheses we wish to explore, but may not support the study of fine-grained cognitive process models, nor, at the other end of the scale, are we attempting to address larger workflow issues.

Our hope is that we will be able to identify some of the conditions under which various resources and tools provide a measurable impact on the translator's productivity or performance. While there are many different kinds of translation tasks, we are interested in broadening the notion of "translation" even more broadly, to incorporate anyone who is attempting to draw information from a foreign language source text and render it within some target language text. This broad definition will include those who wish only to generate a brief summary (or "gist") of a foreign language document, capturing the "salient points" made in the source document. It will include even those who may have extremely narrow "information needs" that are being applied against a source document, such as filling out an information template or questionnaire (e.g., "does this article mention my company's product?", "Does this article talk about cell phone technology?", etc.), or even simply assigning a topic label to a document.

Our reasons for expanding the scope of the analysis are motivated by both practical and theoretical considerations. The theoretical interest is to provide a broader continuum of behaviors which incorporate "foreign language document understanding" and so be better able to tease apart some of the issues surrounding target language composition vs. source language comprehension. The practical interest is that there is an increasingly rich set of ways in which foreign language material is being used within our highly interconnected society. Cross-language information retrieval can result in users attempting to extract meaning from documents in a foreign tongue for which the user is not fully literate. In some organizations relatively junior linguists might be responsible for routing foreign language documents to more experienced translators on the basis of intermediate language skills. Even for the task of full translation, some organizations need to sort through a large amount of

material only some of which may be relevant, requiring the translator to constantly assess the importance and level of effort required to capture the meaning of each sentence within a high-tempo work environment. It is desirable for our empirical analyses to be able to account for this wide range of interactions with foreign language material.

2. Resources as Tools, Tools as Resources

In the context of assessing the value of a resource to the human translator, it is impossible to fully divorce the "static resource" (e.g., a bi-lingual technical dictionary) from the way in which the resource is made available to the translator. Whether it is in form of a hardcopy book, a human mentor, or a computer program, the "interface" between the human and the resource is a key variable that greatly influences the utility of the underlying information. This inevitably creates the opportunity for a confounding influence on the empirical results – a lousy index or lexicographic ordering system can render the richest bilingual dictionary useless to the dictionary reader. Similarly, an annoying or ineffectual user interface in a computerized version of this same resource can thwart the ability of an empirical experiment to identify its underlying value.

These observations indicate how ambiguous is the division between "resource" and "tool" within the scope of experimental analysis of their utility in practice. They also point to a significant caveat that will need to be appended to many empirical results – sometimes a result will provide only a lower bound on the utility of a given resource, since a better interface might enable the resource to be even more useful. We have attempted to provide a relatively consistent framework for interacting with the various tools and resources provided within the experimental platform we have built, so that some of the issues might be said to be "held constant." Nonetheless, the nature of human-computer interaction is extremely complex, so the empirical results that come out of the use of this experimental translation platform must always be viewed with these issues in mind.

3. Resources and Tools to Support Translation

In as many cases as possible we have tried to incorporate strong commercial resources and tools so that the results of our empirical studies reflect as much as possible on the state-of-the-art abilities within these areas. Since we were committed to tight integration and logging, the unfortunate result is that there were many resources (e.g., dictionaries) or applications (e.g., translation memories) for which our desire for integration ruled out many excellent commercial offerings. It is important to note that we are not attempting to evaluate the performance of the individual components that are being integrated, but rather our goal is to perform an evaluation of the relative and absolute value of a class of resource/tool to the translation enterprise. Of course, we can't get around the fact that that we are limited in the

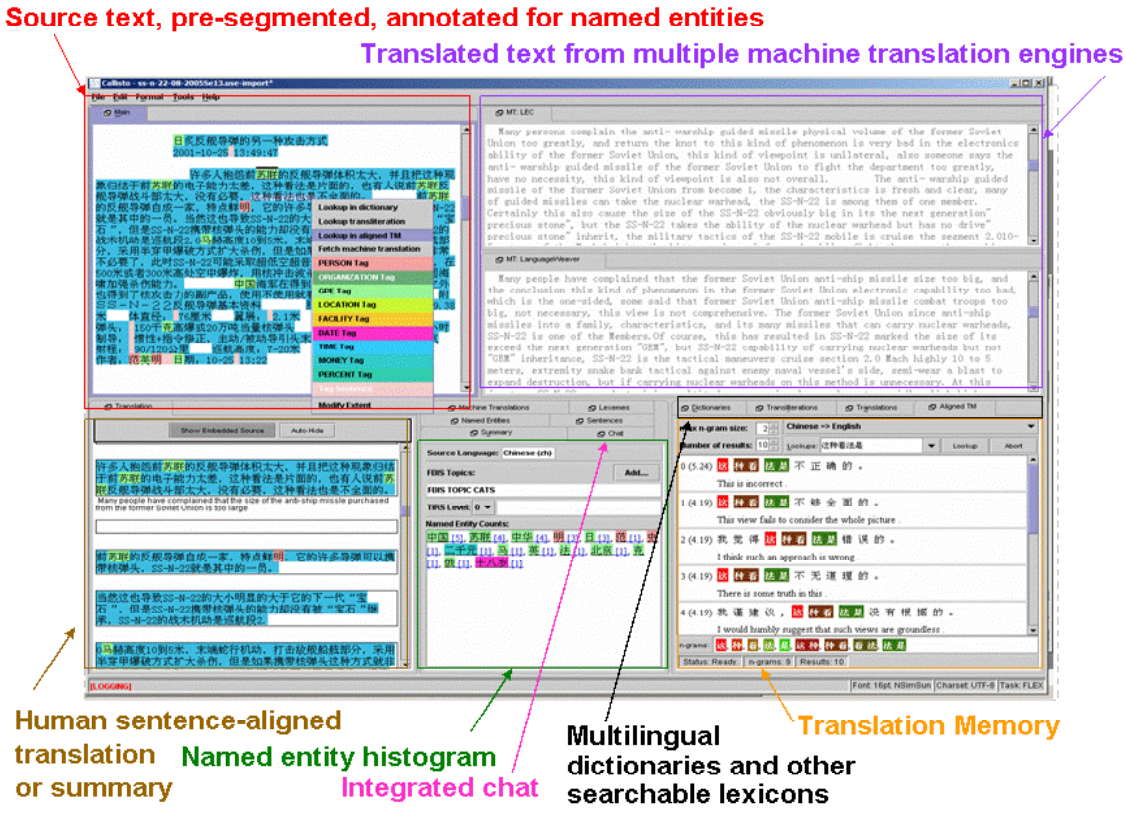

Figure 1: Screen image of the C/Flex translation experiment platform. The user has selected a sentence in the MT pane, with the result that the aligned sentences in the source text pane and the other MT output panes are also aligned and displayed. The dictionary pane displays the results from an earlier query.

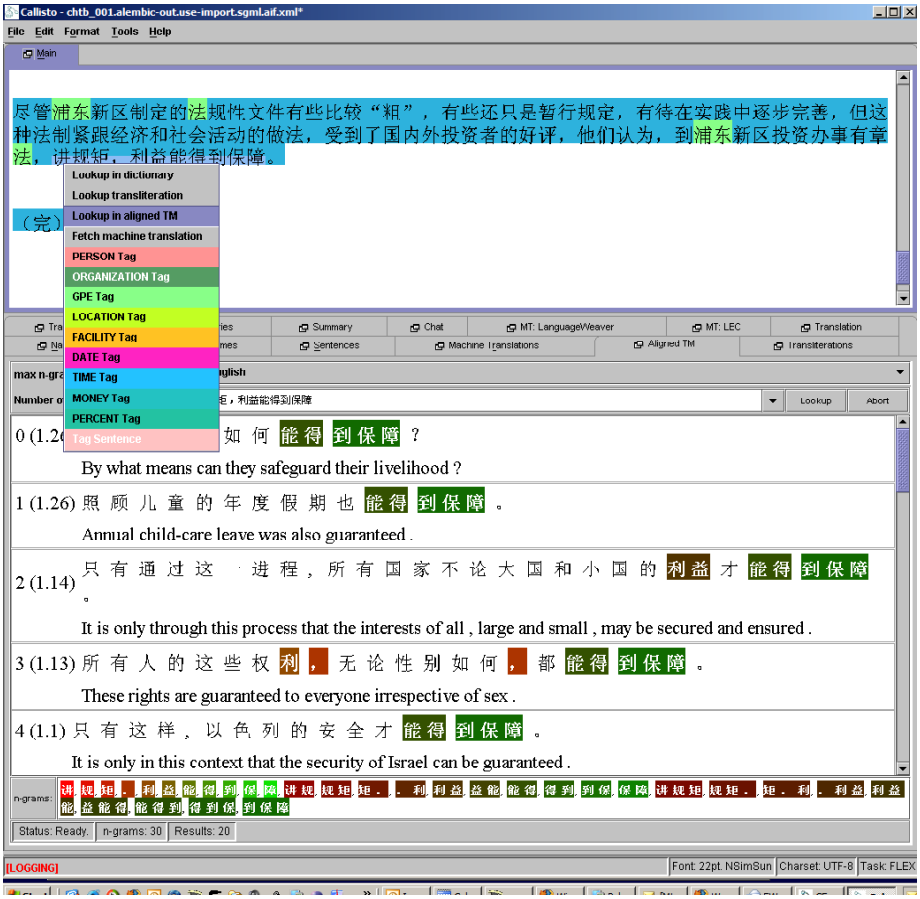

Figure 2: Screen image of a query to and search results from the translation memory (TM).

number and kind of resources and tools we have been able to put together “under one roof,” so the logging data we capture will necessarily reflect the particular resources and tools we happened to integrate.

The experiment platform that we have created is called CalliFlex (and commonly abbreviated as C/Flex). The current version of CalliFlex has been set up for handling three main source languages: Chinese, Arabic and Spanish, though it is easy to add components, resources and tools to support additional languages, if they are available. The target language in all of its configurations has been English. CalliFlex integrates the capabilities and resources listed below.

1. Multiple (usually two) machine translation (MT) systems for each language;
2. An integrated and easily accessible and searchable set of bilingual dictionaries, monolingual dictionaries, onomastica (mono- and bi-lingual person name resources), and other bilingual transliteration and translation resources and automated tools;
3. A large pre-populated translation memory;
4. Source language automatic natural language processing that can identify
   a. Sentence boundaries;
   b. Word boundaries;
   c. Parts-of-speech for each word; and
   d. Named entities (distinguishing among names for persons, organizations, locations, geo-political-entities [“GPEs”] and artifacts);
5. A text chat tool that enables translators to collaborate among CalliFlex clients.
6. A spell-checking facility in the target language translation and/or gist panes (currently restricted to English).

Each of these types of tools/resources are presented to the user within distinct “panes,” and each of these panes can be independently placed anywhere within or outside the borders of the application. The application allows for within-border pane movement in a manner similar to that supported in the Eclipse development environment, enabling an efficient means of filling up the available screen real estate with the selected components. Figure 1 shows an image of one possible layout of the application components

After the user selects a document to import into C/Flex and identifies the language of the source material, the application proceeds to invoke the natural language processing component that identifies sentence boundaries, word boundaries, part-of-speech assignments and named entity expressions. Of course, these automatically derived analyses can and will include errors. (Errors in named entity recognition can be corrected directly by the user -- the system incorporates most of the annotation editing features from the Callisto annotation tool from when it is

derived. We do not yet support editing sentence and word segmentations, or part-of-speech assignments.)

The derived sentence boundaries are then used to invoke the multiple MT engines iteratively, enabling the source sentences and the MT output sentences to all be aligned. The user is able to browse through the document, sentence by sentence, maintaining all the "views" of the sentence in synchrony – source, multiple MT renderings, user's translation.

The CalliFlex prototype also incorporates a pre-populated translation memory (TM). The query interface to this resource generates all possible distinct adjacent multi-word phrases (the user can control the maximum length of these multi-word phrases), and then searches the TM using standard information retrieval metrics for establishing sentence similarity between the collection of phrases and the target source sentences. (In the case of Chinese, these n-grams are measured in characters.) The sentence pairs (source language and target language human translation) returned by the TM search are presented in order of decreasing similarity, and the various multi-word phrases generated as the search query are separately highlighted in the returned sentence pairs. Figure 2 shows a sample of the TM interface when translating a Chinese source document.

The data used to pre-populate the Chinese and Arabic TM data sets is taken from the TIDES 2005 evaluation corpus. These are a mix of general reporting, magazine articles and parliamentary proceedings. The data we will be using in our controlled experiments will come from the general news, so we expect there to be a reasonable intersection of genres between sources articles and TM data.

Queries to any of the resources/tools (dictionaries, name resources, MT modules, transliteration modules, TM) can be invoked either directly from the source text (via word selection and selection from a pop-up menu) or via direct query type-in (assuming the user has access to the appropriate type-in methods on the client computer for that language), and can be invoked in either language direction, source→target or target→source. This ability to change directionality and enter user-generated text allows translators to explore different variations of the source words as well as explore the behavior of the tools/resources themselves under different conditions.

Translation often involves collaboration with other experts, and in some experimental contexts we wish to be able to allow this collaboration to take place while being tracked by the CalliFlex application. For this reason we have incorporated a simple multi-party text chat tool within the client.

The CalliFlex architecture has been developed to enhance the ability for rapid integration of third party tools and resources. Figure 3 illustrates how most of the resources and tools are made accessible via a Tomcat web server, enabling multiple CalliFlex clients to access them via web-based protocols. The dictionaries and transliteration resources are stored in the OLIF2 interchange format (McCormick, 2004), which are then indexed by a Lucene search engine, enable full-text and fielded search from the client.

4. Application Logging and Post-Experiment Analysis

The CalliFlex prototype captures a log of all of the following types of information, associated with each user session. Every log identifies the user ID (possibly an anonymous but unique ID), and each entry includes a time stamp.

1. Start and end times – when a document is first imported into the tool, the source document's language, the target language of the translation, the various features within the CalliFlex tool that have been made available to the user, and when the state of the system (including translation/summary) is saved or exported at the end of a session.
2. Resource lookup – When the user queries a resource such as a mono- or bi-lingual dictionary, transliteration resource, translation memory, etc. This also includes various tools that perform automatic processing on the query string, such as machine translation engines (used as a dictionary), transliteration algorithms, etc. The logged information includes whether the string was entered directly by the user or whether it was copied (swiped) from one of the application panes, in which case the identity of the source pane will

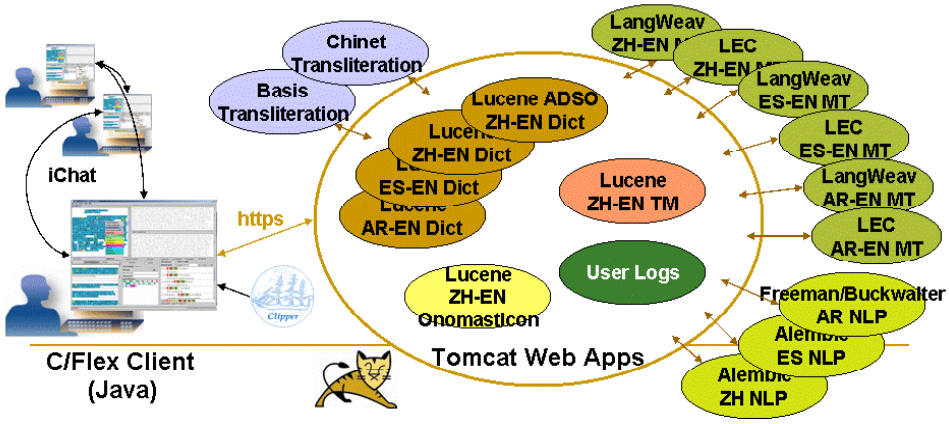

Figure 3: The Calliflex (C/Flex) Architecture. The resources and tools can be hosted locally or remotely, with remote access via TCP/IP web app. Simple application wrapper and protocols encourage rapid incorporation of new tools and resources.

also be included.

3. Translation pane updates – The source of text entered into the translation/summary pane is identified, whether it is from a copy-and-paste (e.g., from one of the MT output panes) or typed in by the user.
4. Annotation edits – As noted earlier, the source text is automatically processed to reveal predicted sentence boundaries, word-boundaries, part-of-speech assignments and named entities. Each of these types of assignments are prone to error (varying greatly depending on the genre of the source text as well as the extent to which it is complete sentences vs. abbreviated text as one often sees in web sites). The tool currently supports user editing of named entities and sentence segments, but not of the word segments and part-of-speech assignments.

As noted in the introductory section, our experimental focus is currently on the assessment of the relative contributions made by state-of-the-art computational aids such as machine translation, automatic named entity recognition, name transliteration, etc. This focus has meant that some of the more detailed levels of logging are left out that are present in other tools. For example, Translog (Jakobsen, 1999) is a powerful text editing analysis tool that has been used successfully in studies of translation (as well as other text editing tasks) that provides character-by-character update timing. We have not attempted to replicate this kind of fine-grained logging for the present experimental goals.

However, given the large number of different sources from which a given piece of target (translation) text can derive from, this is an addition and important piece of information that we are able to track. Thus, the logs will keep track of whether the text that is entered into the translation was copied from a transliteration pane, a machine translation output pane, a TM sentence, etc., and a timestamp on when this update happened.[1] We will also track deletions from the translation pane.

We have only recently brought the CalliFlex tool to a state sufficient to support experiments, so as of the date of writing this paper we do not yet have sufficient empirical experiment results to analyze. At this point we can only report on our plans for testing various hypotheses and the experimental setup and data we intend to capture to explore these hypotheses.

4.1. Translation/Summarization Productivity

One of the first questions on which we will concentrate in our studies is that of translator productivity: Can the tools brought together within the CalliFlex experiment platform support measurable productivity gains in full translation or target language summaries without any diminution in quality? Our hypothesis is that this is indeed the case for relatively junior translators and for those working in highly dynamic/technical subject disciplines. As translator skill increases, the utility for resources and tools diminish. A corollary hypothesis is that resource/tool utility in these target populations is increased in relation to the amount of time pressure imposed on productivity. While these hypotheses are modest, we hope to provide concrete empirical evidence, and also begin the process of identifying the relative contributions of different kinds of resources and tools within these different translator populations and work contexts.

External time pressures are one of the dominant aspects of the translation and summarization work contexts that we wish to study. Earlier work on time pressure in translation (e.g., Jensen, 1999) has attempted to elucidate cognitive explanations for the differences in behavior observed under different temporal constraints and with translators of different skill levels. Our data collection and analysis will concentrate initially on the differential influence that automated methods and enhanced resources play in both the translators' use of these tools, and to the extent possible the degree to which these tools actually do ameliorate the deleterious effects of tight time constraints. As in the earlier work, we anticipate, and will attempt to carefully track, the different performance characteristics that will be associated with highly experienced translators versus more junior translators.

Our experiments will be conducted in the following manner:

1. Each subject will be given a questionnaire, in which we inquire about their level of expertise in the language, translations skills and experience, the kinds of tasks they perform in their normal work, etc.
2. CalliFlex Tool Suite Training. This will be an important variable to measure. The tool is fairly complex and provides a great deal of user-customization options.
3. Summarization/Translation. The subject will be provided a fixed number of documents and will be asked to translate, or in a different experimental context be asked to provide a summary translation, of each document's contents. In the case of the summary, a specific expected length will be determined. The summary will be "domain independent" – the subject should attempt to capture as much of the "important" elements of the document as possible.
4. The subject will be provided these materials in four fixed time-period segments. Within each segment there will be one of two experiment conditions adopted: enabling all of the CalliFlex tools and resources to be available to the subject, or enabling only the ability view the source and write the translation/summary (with all other resources available only via hardcopy documents). The order in which these conditions are presented will vary among subjects.
5. A post-experiment questionnaire will be given to the subject, in which we ask a variety of subjective assessments of the tools and resources provided within the experiment, ideas for improvements, how well the experiment seemed to capture their usual work environment, etc.

[1] Note that if a user elects to type in some text from one of these other panes, as opposed to the more natural copy-and-paste action, the tool will not be able to determine the actual source of the text.

6. Post experiment analysis. The data will be analyzed from a number of perspectives. A particularly important analysis will be assessing the quality of the translation or summary, and associating this quality against the experiment conditions (with or without various tools) and the amount of time it took for the subject to generate this product. Our test data include eleven human translations for each source document. We will use various simple domain-independent Likert numerical quality scales (1 – 5) by which multiple evaluators will grade the quality of the translations/summaries. While such an evaluation metric may be crude, our focus is on measuring performance differences in relatively junior translators, where there is often a fairly high variability in translation or summary quality. In the case of full translations, we will also make use of the standard MT evaluation metrics such as BLEU (Papineni, et al, 2002), though these have limited discriminatory power. In the case of summaries, we will employ some of the techniques developed in the Document Understanding Conferences (Dang, 2005) evaluations to establish how many of the key elements of information have been included in the summary. These key elements are identified by comparing against summaries generated by multiple evaluators. These evaluators will work against the translated documents rather than the source documents.

Our "standard" experiment protocols do not presently call for any formal role for Think Aloud Protocols or TAPs (Lörscher, 1991). This is mostly because we wish to attempt to reduce the per-subject costs of the experiment sufficiently to enable a relatively large subject population, and thereby increase the opportunity for statistically significant numbers in our captured data. However, we are cognizant of the immense influence that particular user interface design elements can have on our experiments. For this reason we intend to conduct small scale, mostly in-house interactive experiments in which we record the video and audio of the experiment session, and in which we may introduce questions and ask for feedback. These sessions will not attempt to rigorously pursue the TAP methodology, however.

Due to the complexity of the tool, we have developed a fairly rigorous training and exercise regimen to familiarize subjects with the wide variety of tools and information sources. These training sessions include many opportunities to provide anecdotal feedback to us on what they think about the tool's components and their utility in performing translations or summarizations.

The discussion of the experiment platform and experiments has so far concentrated on the translation and "gisting" (summarization) tasks. If we are able to obtain sufficient experimental subjects and associated experiment materials we hope to explore a number of similar hypotheses associated with different translation "tasks" such as name finding, template filling, reading comprehension tests, etc., as well as establishing different kinds of experiment conditions in order to mimic those that might be found in different work environments – for example, subjects occupying a crowded room in which others are performing similar tasks; allowing collaboration among subjects with similar or different levels of skill, etc.

5. Current Status and Plans

We are just now, in the Spring of 2006, beginning to perform experiments with as many subjects as we are able to find. We intend to make the CalliFlex application available to other researchers without charge in order to encourage a greater investment in the empirical study of translation and how emerging linguistic resources and tools can enhance the productivity and quality of this important activity. The application will enable others to incorporate different resources and computing capabilities that may open up new experiments and test different hypotheses.

6. References

Dang. Hoa Trang (2005). Overview of DUC 2005. Presented at the Document Understanding Workshop, HLT/EMNLP Conference, October 9-10. Vancouver, B.C., Canada (http://duc.nist.gov/).

Danks, J., G. M. Shreve, S. B. Fountain, M. McBeath , eds., (1997). Cognitive processes in translation and interpreting. London: Sage Publications.

Jakobsen, A. L. (1999). Logging target text production with Translog. In G. Hansen (ed.), in Probing the process in translation: methods and results (Copenhagen Studies in Language 24). Copenhagen: Samfundslitteratur, p.9-20.

Jensen, A. (1999). Time Pressure in Translation. In G. Hansen (Ed.), *Probing the Process in Translation: Methods and Results* (pp.103-119). Denmark: Samfundslitteratur.

Lörscher, W. (1991). "Thinking-aloud as a method for collecting data on translation process". In S. Tirkkonen-Condit (ed.). Empirical research in translation and intercultural studies. Tübingen: Gunter Narr, p.67-77.

McCormick, Susan (2004). Using OLIF, The Open Lexicon Interchange Format. Presented at the 6th Conference of the Association for Machine Translation in the Americas, Georgetown University, Washington DC, September 28 - October 2.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL), Philadelphia, July, pp. 311-318.

Tirkkonen-Condit, S. (1986). Empirical studies in translation: textlinguistic and psycholinguistic perspectives. Joensuu: University of Joensuu.