The Bilingual Web Dictionary on Demand

Henrik Selsøe Sørensen
Copenhagen Business School
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
hss.fr@cbs.dk

Abstract

What do you do when you need to find terminology in a foreign language and available bilingual sources are of no help at all? With the fast-growing complexity in all fields of knowledge and the parallel creation of neologisms needed to distribute the new knowledge, dictionaries and term databases are lagging behind, in particular when you work with two less widely spoken languages. This paper investigates methods and strategies for solving the problem and proposes to lay the grounds for a Bilingual Web Dictionary on Demand. This virtual dictionary is conceived as a number of knowledge-based methodologies allowing users to get across the language barrier using standard search engines and the Web as corpus. The key methodology aims at identifying a target language text containing an equivalent to the source language term using a so-called Cluster Method. Once a candidate term in the target language is identified, it must be validated. Intended users are translators, lexicographers, terminologists who must be bilingual in order to be able to select and adjust clusters.

1 Introduction

With a fast-growing complexity in all fields of knowledge and constant creation of neologisms, static dictionaries - printed or electronic - usually offer no or little help to translators, lexicographers and terminologists. New phenomena appear and are named daily, but it takes time before they find their place in mono- or multilingual resources, in particular when it comes to less widely spoken languages. What is the best practice for overcoming the language barrier when searching for target language (TL) equivalents to source language (SL) neologisms?

Translators already make extended use of search engines and texts on the Web for checking terms, term usage and collocations, etc. (Varantola 2000). But if they do not have any clue as to what the TL term is, they have a problem. They may first try conventional resources like static dictionaries, bilingual electronic repositories, term databases, e.g. Eurodicautom, and any multilingual dictionaries, glossaries, vocabularies and word lists available on the Web. If relevant parallel corpora are available, term extraction from these may prove useful. In some cases there are ways to find a useful monolingual TL resource, e.g. an ontology, a thesaurus, an encyclopaedia or a newspapers whether printed or electronic or in any other form. Asking a friend or colleague is also a good strategy. All of these conentional strategies should be tried, but if everything fails, the only strategy left is to go hunting, i.e. do terminology mining in an unaligned and non-parallel corpus, in particular in the biggest, most fascinating and definitely most updated of them all - the Wild Web. Whether all of the texts on the Web qualify as a corpus or not is of little interest in this context, no corpus can be said to be truly representative anyway, as argued by Kilgarriff and Grefenstette (2003). The Bilingual Web Dictionary on Demand project is an attempt to define, discuss and group best practices for hunting TL equivalents, a first step on the way to the ultimate Web-based dynamic dictionary on demand.

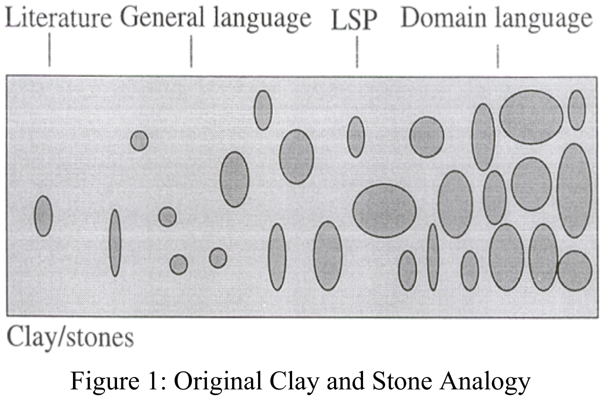

Figure 1: Original Clay and Stone Analogy
(Melby, A. & Warner C.T., 1995)

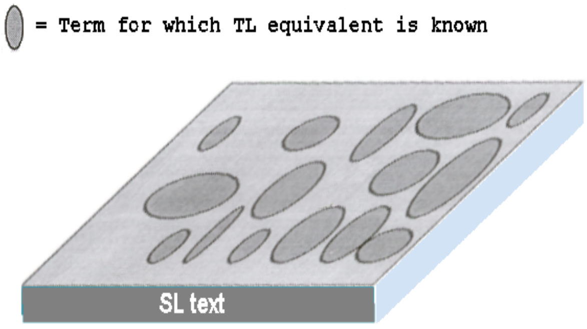

Figure 2: Clay and Stones of a Text to be Translated

2 Hunting for TL Equivalents

Melby, A. & Warner C.T. (1995) described words as “chunks of pliable clay, terms as hard stone”, cf. figure 1, in an inspiring attempt to characterize general language and LSP (language for specific purposes).

This analogy will be used here in a slightly different version: The slice of clay and stones in our context represents a single complete text as suggested by figure 2. ‘Term’ and ‘concept’ will be used as synonyms in this paper, ‘concept’ being preferred when ‘concept systems’ are discussed later on.

In the case of a straightforward translation where all equivalents are known or have been found by the translator, the stones are shown in grey as in figure 2.

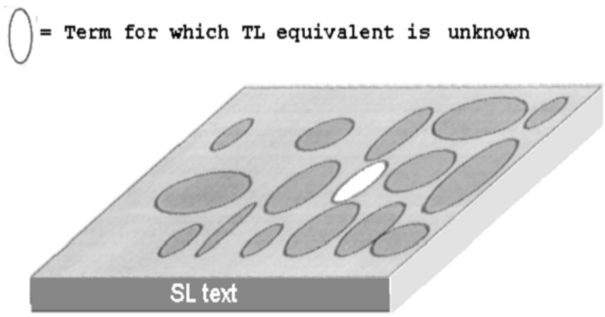

Figure 3: TL Equivalent to "White SL Term"

If for a SL term no equivalent TL term is known or can be identified, the situation would be pictured as in figure 3, the problematic term being marked as a "white stone". It is time to test the Cluster Method.

3 The Cluster Method

As step 1 select three salient terms from the SL text for which the TL equivalents are well documented. These equivalents will be used for a cluster search.

Step 2 is to use the three TL equivalents for a Google search with all of the words in order to possibly find comparable texts on the Web. The hypothesis is that if a cluster of salient terms leads to a TL text containing the same cluster, then it is possible that this text also contains the unknown TL equivalent. The planning phase of a search for a cluster of three is illustrated in figure 4.

If too many texts are found with this cluster of terms, additional salient terms should be added. If no or very few texts come up, the cluster may be reduced, or one or two cluster terms may be replaced. In this way, the user must go on hunting for a contentwise comparable TL text.

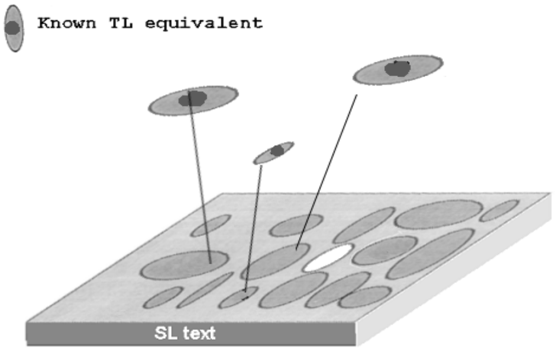

Figure 4: Three TL Equivalents Selected for a Cluster Search

Step 3 consists in analysing TL texts found by Google in search for the unknown TL equivalent to the "white stone". To do this, corpus analysis tools may prove useful, but the user must also do some reading and use his knowledge. If the method succeeds, at least one of the TL texts identified by the Cluster Method contains a candidate TL equivalent as pictured in figure 5. If a suitable TL text could not be found on the first try, the cluster should be modified using e.g. more than three salient terms from the text.

Once a candidate TL term has been located, it will of course have to be validated in printed or electronic resources and through new Web searches. If the SL term is a neologism, it is likely that the TL term is, too.

This method will only work if there is no conceptual gap between SL and TL: The method obviously cannot identify an equivalent which does not exist. In those cases, however, the method might help shed light on the discrepancies, cf. the Ontology Enhanced Cluster Method proposed below.

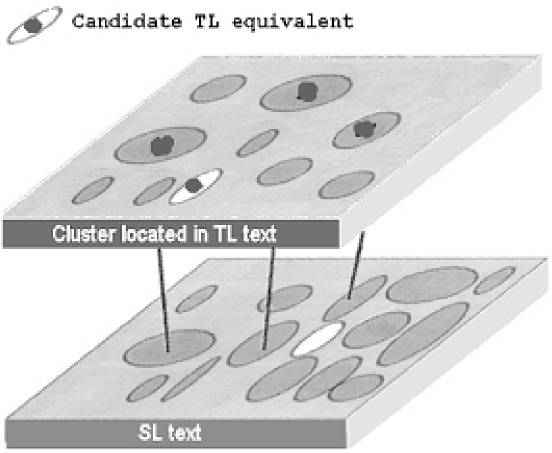

Figure 5: TL Equivalent to the "White SL Term" Succesfully Located through Cluster Search

3.1 Example: Translation French – English

A French text to be translated into English (Conférence mondiale contre le racisme, 2001) contains the term <u>servitude pour dettes</u> for which no TL could be found, cf. this extract:

"*Comment fonctionne le piège*
Les trafiquants ont recours à toutes sortes de méthodes de recrutement et n'hésitent pas à enlever purement et simplement leurs victimes ou à les acheter à leur famille. [..] Son entrée ou son séjour dans le pays de destination est généralement illégal, ce qui la met dans une situation de dépendance accrue à l'égard des trafiquants. Le système de la <u>servitude pour dettes</u> est largement utilisé. Il permet de contrôler les victimes et de tirer indéfiniment profit de leur travail."

Four salient terms were selected in this case for a cluster search because they are considered representative of the text and because their TL equivalents were known: 'smugglers', 'debt', 'destination' and 'victims', cf. the SL terms which are boldfaced in the text below.

"*Comment fonctionne le piège*
Les trafiquants ont recours à toutes sortes de méthodes de recrutement et n'hésitent pas à enlever purement et simplement leurs ***victimes*** *ou à les acheter à leur famille.*

[..] Son entrée ou son séjour dans le pays de ***destination*** *est généralement illégal, ce qui la met dans une situation de dépendance accrue à l'égard des* ***trafiquants****. Le système de la servitude pour* ***dettes*** *est largement utilisé. Il permet de contrôler les victimes et de tirer indéfiniment profit de leur travail.*"

For this cluster of four, Google found around 1.600 texts. The page ranked as number 1 happened to be a report called "Slavery, Abduction and Forced Servitude in Sudan" (Bureau of African Affairs, 2002), and so the first candiate TL equivalent had already been identified. Given the huge number of texts returned, it is recommended in any case to expand the cluster in order to narrow down the number of texts and thereby possibly increase the relevance of the located texts. When a fifth TL term, 'countries', was added to the cluster, the following text came up in the top five (Xatrix Security: IT news, 2003). The cluster is boldfaced, the candidate TL equivalent is underlined.

"*Although the Geneva-based ILO gave no overall figures, it said the United States was believed to be the* ***destination*** *for 50,000 trafficked women and children each year alone with New York and California the main entry points. The report said slavery was increasingly rare but still found in a few* ***countries****. [..] "****Victims*** *frequently find themselves trapped in* ***debt*** *bondage and other slavery-like conditions." [..] Authorities had difficulty detecting the trade as it is often carried out by international gangs who find it less dangerous than drug smuggling. People* ***smugglers*** *have rarely been caught and the punishments handed down were usually lighter than for drug smuggling, the ILO said.*"

After further hunting, a total of four candidate TL equivalents were identified, i.e. 'forced servitude', 'debt servitude', 'debt bondage' and 'debt slavery'. The validation procedure which plays an important role in the strategy eventually seemed to nominate 'forced servitude' as the most reliable equivalent to 'servitude pour dettes' based on frequency and reliability of the sources.

4. The Ontology Enhanced Cluster Method

Many new phenomena which are discussed worldwide, are perceived independently of national barriers, but are of course named in each language. For such cases, the Cluster Method is likely to succeed. However, far from all phenomena are relevant across frontiers, many appear and are defined on a national level, and these are not likely to be discussed and named outside the country where they were created. In such cases of conceptual gaps, the Cluster Method probably will not work. Nevertheless, lexicographers and terminologists still may have to bridge the gap and come up with translations.

Such gaps occur for concepts at a certain level in a concept system. When moving upwards in that concept system toward superordinate and broader concepts, the likelihood of finding equivalent concepts outside the particular country increases. In the same way, moving downward towards subordinate concepts of that system, more gaps are likely to occur. Although a certain disease is defined locally in a way differing from how it is defined abroad, it is likely that the type of disease to which it belongs, i.e. its superordinate concept, has an equivalent outside the country in question. It is much less likely that a subtype of that disease would have an equivalent.

For "stones" representing conceptual gaps, the Ontology Enhanced Cluster Method proposes to try to identify their superordinate concepts in the TL and then use clusters consisting of TL terms for these concepts to make the search, cf. figure 6. The search result would not be a text with direct TL equivalents, but in the best of cases a relevant text with material providing the translator or the terminologist with sufficient knowledge to render the local SL concept in the TL.

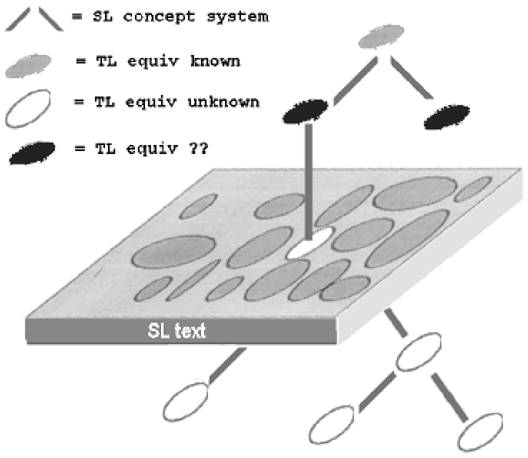

Figure 6: A Superordinate Concept (Grey) to be Used for a Cluster Search

The steps would be the following:

1. Find and explore an existing or potential SL ontology containing the locally defined concept.
2. Move upwards in the hierarchy from that concept and select the first superordinate concept which is known and defined in the TL.
3. Alternatively, use characteristic subtrees from the SL ontology in a cluster search in order to identify relevant TL ontologies as shown in example below.
4. Explore the identified TL ontologies in search for relevant concepts that may document what the TL reality is like and ultimately help rendering the problematic SL concept in the TL.

The Ontology Enhanced Cluster method is demanding to use and at a very experimental stage, but may help squeeze out more data, cf. the example below.

4.1 Example: Translation Danish – English

The Danish word 'museskade' literally means 'mouse damage' and is defined as pains in the arm, elbow and neck known by users of a computer mouse. When a translation of 'museskade' was needed in 2002, no help could be found anywhere, and the original Cluster Method proved unsuccessful, apparently there was a conceptual gap. Obviously, however, it is not only mouse users in Denmark who suffer from these symptoms, so the disease had to be known abroad. A first try with a literal translation did not work, 'mouse damage' and 'mouse disease' both only led to texts about living mice and their ups and downs. It was also not possible to find a Danish ontology or system of concepts taking this disease into account. However, an ontology-like description of muscular diseases in Danish was found which linked various diseases to certain symptoms. 'Muscular disease' seemed to be the superordinate concept that would serve as link between SL and TL. A cluster search not with terms from a SL text but with symptoms and body parts from the ontology was performed. This ontology generated cluster search contained: 'stiffness', 'numbness', 'burning pain', 'forearms', 'hands', 'wrists'. It actually lead to an appropriate ontology-like TL description (Moore, K.L., 2004) that also revealed a candidate term in English. In fact, again several candidate terms were found after som further hunting:

- computer related RSI (RSI: 'repetitive strain injury')
- computer related CTD (CTD: 'cumulative trauma disorders')
- 'Mouse injury' which would be stylistically equivalent to the Danish term also came up, but it had only a few dozen occurrences on the web when the meaning 'injured mouse' was disconsidered <18.11.2003>.

Whether there is actually full equivalence between Danish and English is for a medical specialist to decide. In any case, there is a clear stylistic discrepancy, the Danish term being straightforward and very frequent in general language, whereas the two frequent English terms are highly technical.

5. A Bilingual Web Dictionary on Demand

The Cluster Method and its variant The Ontology Enhanced Cluster Method should be seen as initial contributions to a Bilingual Web Dictionary on Demand: the WebDoD. The goal of the WebDoD project is to create a new type of dynamic dictionary that draws directly on texts available on the web and extracts bilingual data on demand. There is obviously a long way to go. At least in its pioneer phase, the WebDoD requires quite skilful users, but further developments should aim at helping users perform their part of the task more easily using a package of sophisticated tools and methods. The tools in question are in fact to a certain extent already available, as will be seen from the "want list". What is particularly needed at this stage is a seamless integration of the items listed below:

1. Easy access to static bilingual resources for conventional searches.
2. Cluster methods whenever the static resources fall short.
3. Selection of SL cluster words using sophisticated corpus analysis tools, e.g. System Quirk, Intex, WordSmith Tools.
4. Search engines.
5. Statistical language processing methods as suggested by Fung, P. & McKeown, K. (1997).
6. In-depth analyses of selected TL texts, in partcular System Quirk tools like KonText and Ferret.
7. Ontological query methods as developed by the OntoQuery Project (Madsen, B. N. et al., 2001).
8. Validation practices.
9. Storage and recycling of validated results among WebDoD users.

Bibliographical references

Web sites visited 19.2.2004.

- Bureau of African Affairs (2002). Slavery, Abduction and Forced Servitude in Sudan. http://www.state.gov/p/af/rls/rpt/10445.htm#glossary
- Conférence mondiale contre le racisme (2001). Dimension raciale de la traite des personnes, en particulier des femmes et des enfants. http://www.un.org/french/WCAR/e-kit/issues.htm
- Eurodicautom, Multilingual Term Bank of the European Commission. http://europa.eu.int/eurodicautom/Controller
- Fung, P. & McKeown, K. (1997). Finding Terminology Translations from Non-parallel Corpora In Proceedings of the 5th Annual Workshop on Very Large Corpora (pp. 192-202). Hong Kong. http://citeseer.nj.nec.com/fung97finding.html
- Google - Searching 4,285,199,774 web pages. http://www.google.com/
- Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. In Computational Linguistics, Special Issue Vol. 29 Number: 3 (pp. 459-484).
- Intex Linguistic Development Environment. http://www.nyu.edu/pages/linguistics/intex/
- Madsen, B. N. et al. (2001). Defining Semantic relations for OntoQuery. In P. A. Jensen, & P. Skadhauge (Eds.), Ontology Based Interpretation of Noun Phrases (pp. 57-88). Syddansk Universitet , , link from: http://www.ontoquery.dk/publications/index.php
- Melby, A. & Warner C.T. (1995). The possibility of language - a discussion of the nature of language, with implications for human and machine translation. Benjamins Translation Library, Benjamins, Amsterdam.
- Moore, K. L. (2004): Computer Related Repetitive Strain Injury Site http://eeshop.unl.edu/rsi.html
- System Quirk Language Engineering Workbench http://www.computing.surrey.ac.uk/SystemQ/tracker/
- Varantola, Krista (2000): Translators and disposable corpora. In Proceedings of Corpus Use and Learning to Translate (CULT), Bertinoro, Italy.
- WordSmith Tools http://www.lexically.net/wordsmith
- Xatrix Security: IT news (2003): http://www.xatrix.org/article330.html .