Terminological Ontology and Cognitive Processes in Translation*

Fumiko Kano Glückstad

Dept. of Intl. Language Studies and Computational Linguistics, Copenhagen Business School
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
fkg.isv@cbs.dk

Abstract. This work is based on a terminological ontology method proposed by Madsen *et al.* (2004; 2005). The main purpose is to bridge a semantic relation between domain specific terms in two languages by mapping two language-dependent terminological ontologies. To explain why this method is preferable from the view of the cognitive process in translation, the terminological ontology method is contrasted with a study made by De Groot (1997) in the area of bilingual memory research.

Keywords: terminology, ontology, bilingual memory, ontology matching, translation

1 Introduction

Imagine a situation where non-English speaking European and Asian are debating in English about the issue of the academic degree systems in their respective countries. While a German might be explaining about the Doctor of Science (Habilitation) as their highest academic degree, a Japanese might be having the highest academic degree in Japan - Ph.D. level - in his mind. This imagined conversation shows a typical situation revealing a deep inherent misconception between the two parties since each of them has their own conceptual - and correct - understanding of the highest obtainable academic degree in their respective countries.

Statistical Machine Translation (SMT) is one of the most widespread machine translation approaches. However, when considering rare language-combinations, the lack of "direct" parallel corpora is a critical concern for successfully applying the SMT approach. In case of the rare language-combination: Danish and Japanese, it is possible to collect a reasonably sufficient amount of corpora between Danish and English as well as English and Japanese. However, the availability of bilingual corpora that directly link Danish and Japanese are very limited. Hence, a transitive translation technique using English as a pivot language is often employed in translation of rare language-combinations. A question is how well the transitive machine translation approaches can convey original conceptual meanings of Source Language (SL) words into Target Language (TL) translations in the transitive translation. When considering human translations between rare combinations of languages, human translators typically look up SL-English dictionaries to identify meanings of SL words and then use English-TL dictionaries to determine appropriate TL translations. In such situations, the problem of word sense ambiguation becomes critical. Since a lexical representation often carries several meanings, the transitive translation approach using English as pivot simply amplifies the probability of selecting an inappropriate sense in a TL. Thus the problem of polysemy becomes especially serious in the process of transitive translations.

* Acknowledgement: I would like to thank my supervisor, Hanne Erdman Thomsen, for helpful advices on my project.

This present work is primarily based on a terminological ontology method (Madsen *et al.* 2004; 2005), a domain-specific ontology which is used in standardized terminology works such as ISO 704 (2000). Since the traditional study of terminology is widely described as the standardization of terminology, a concept is ideally referred to by a single term. In other words, one single term unambigously designates one single concept. Thus, the use of a terminological method is highly suitable for domain specific translations. The terminological ontology method is potentially useful for providing a precise map of the terms covering the concepts and their interrelations, as well as to contribute to the identification of equivalences between selected languages. By mapping the language-dependent terminological ontologies between selected languages, it will potentially be possible to identify an appropriate translation in a TL based on semantic-rich information. The Feature Specification (FS) approach used in the terminological ontology method has similarities to the Distributed Conceptual Feature (DCF) model, one of the bilingual memory models proposed by De Groot (1997). The similarities explain why the terminological ontology method is potentially useful for conveying original conceptual meanings of SL words into TL translations.

To outline the scope of this work, Chapter 2 addresses De Groot´s DCF model based on the theory of bilingual memory, followed by the terminological ontology principles in Chapter 3. In Chapter 4, an example of the terminological ontology mapping that has been manually carried out is demonstrated. Chapter 5 discusses issues concerning the mapping algorithms followed by a consideration of evaluation methods. Finally, the conclusions follow in Chapter 6.

2 Distributed Conceptual Feature Model

Sager (1990) contrasts a phenomenon that is widely known within translation theories as follows: *when translating informative source texts, e.g. a manual or a textbook, and if there is no foreign language term that corresponds to the source definition, a translation equivalent is created. The validity of the translation equivalent is restricted to the context where it has been created until it becomes fully accepted as a term with its own definition. When the translation equivalence is fully accepted, it has become associated to a concept in a conceptual system that is expressed in the foreign language*. This phenomenon is indeed well explained by the series of bilingual memory representation models. These consist of two layers: lexical representations and conceptual representations (Potter *et al.*, 1984). The "word-association" model includes a direct connection between the lexical representations of the two languages, while only the SL lexical representation is connected to the layer of conceptual representations (Figure 1-a). On the contrary, in the "concept-mediation" model (Figure 1-b), the lexical representations of the two languages are connected through the layer of conceptual representation that is shared between them. Kroll (1994) has pointed out that the strength of connections among representations changes depending on the proficiency of the TL (Figure 1-c).

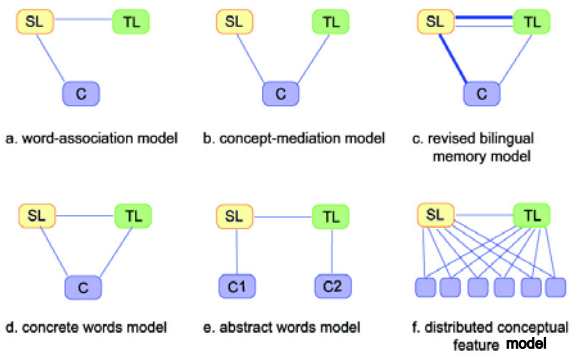

Figure 1: Bilingual memory models.

De Groot (1997) demonstrates effects of word concreteness on the translation performance through her psycholinguistic experiments. She redefines the bilingual memory models in order to explain the word concreteness effects. Concrete words enable two routes to the translation response: one along the direct connection between the lexical representations and the other, indirectly, via the conceptual representation (Figure 1-d). On the other hand, abstract words only permit translation along one route and that is the direct connection between the lexical representations both of which have their own language-specific conceptual representation (Figure 1-e). She elaborates on her theory that concrete words share more of the individual elements of conceptual features (DCFs) between the two lexical representations (in SL and TL) than abstract words do. Hence, the connections between an SL word and its translation into TL are strengthened via DCFs that are shared between them in the case of concrete words (Figure 1-f).

Based on the DCFs model, it can be assumed that, if there is a system which can enlist conceptual features distributed between the two lexical symbols in question, such a system may help translators to perform effective translations in situations where they cannot identify any candidate translation in a TL from very limited language resources. Thus the terminological ontology that can systematically extract conceptual features can be used for the purpose of domain specific translations.

3 Terminological Ontology Principles

The backbone of terminological ontology is comprised of characteristics modeled as typed feature specifications (Carpenter, 1992). Feature specifications (FSs) consist of a feature dimension and a value, i.e. attribute-value pairs. Thus, a representation of a whole concept is a feature structure, i.e. a set of FSs corresponding to the unique set of characteristics, which constitutes that concept.

The terminological ontology is based on several principles. Most importantly, a concept automatically inherits all FSs of its superordinate concepts. For any given concept, a feature of a non-inherited characteristic may be a subdivision criterion for a superordinate concept. This clarification makes it much easier to identify subdivision criteria and for differentiating characteristics in practical terminology work. This means that the same feature attribute must always occur on sister concepts and their value can only appear on one of these sister concepts. Hence, a feature dimension (attribute) that comprises primary FSs must be chosen in such a way that each daughter concept has one and only one FS (value) specified in its mother concept. Hence, a concept must be distinguished from each of its nearest superordinate concepts as well as from each of its sister concepts by at least one FS (Madsen *et al.*, 2004; 2005).

If terms and their conceptual features can systematically be enlisted in this way, it can potentially convey original conceptual meanings of SL terms into TL translations based on the hypothesis that FSs can play a role corresponding to DCFs. To achieve this, it is necessary to obtain a map between conceptual features of SL terms and of potential TL translation candidates. To experiment this hypothesis, two cultural dependent terminological ontologies that represent each conceptual system of SL and TL are semi-automatically developed and manually mapped in Chapter 4.

4 Experiment: Terminological Ontology Mapping (Manual Mapping)

4.1 Method

Corpora

Corpora describing the Japanese educational system have been identified from the "Multilingual Living Information[1]" site provided by the Council of Local Authorities for

[1] http://www.clair.or.jp/tagengorev/en/j/index.html

International Relations and from a pamphlet entitled "Higher Education in Japan[2]" published by the Japanese Ministry of Education, Culture, Sports, Science and Technology. For the Danish educational system, a web-site[3] published by the Danish Agency for International Education has been used as corpora. All corpora are officially translated in English so that it is possible to identify original expressions in SL from documents published by the respective authorities. Thus, all English translated terms and their original expressions are considered as official terms provided from reliable authorities for each country.

Ontology Development

The terms and their definitions describing the educational systems in each country are manually extracted from the English corpora. Based on these terms and their definitions, ontologies representing the educational system in each of the two countries are developed according to the aforementioned terminological ontology principle (Figure 2).

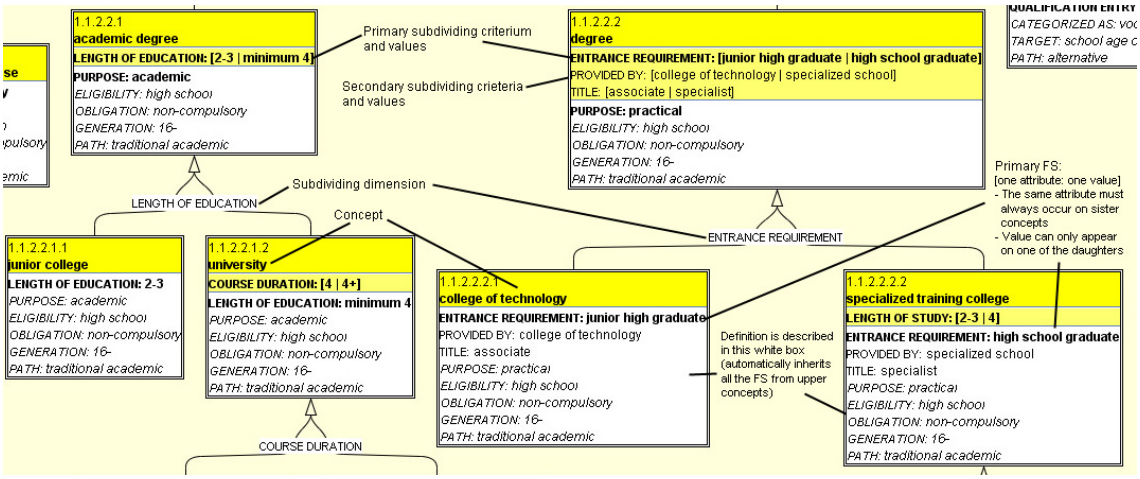

Figure 2: Terminological ontology – Japanese higher education

Ontology Mapping

The basic techniques used in the manual mapping procedure are the string-based matching, the graph-based matching, and the internal structure (feature structure)-based matching (Euzenat and Shvaiko, 2007). All of these three matching techniques are combined as shown in Figure 3.

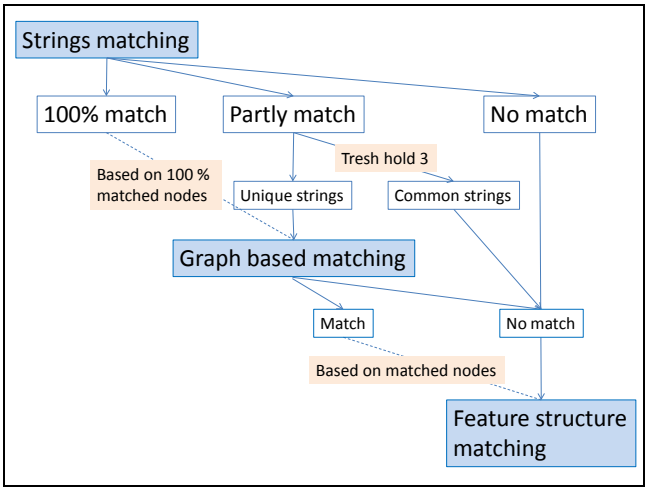

Figure 3: Ontology mapping algorithm

[2] http://www.mext.go.jp/english/koutou/detail/1287370.htm
[3] http://en.iu.dk/education-in-denmark/the-danish-education-system

1) String-based matching at the Lexical Representation (LR) level

The simplest technique is to match strings at the LR level. The expected results are either 100% matching, partly-matching or no-matching. In case of the 100% matching, the terms in the two ontologies are considered as equivalent. For example, the Danish term "videregående uddannelse" is associated with the Japanese term "高等教育 *koto kyoiku*" via their English translation "higher education". This relation indicates that these two terms form a "word-association" of the DCF model described in Chapter 3.

The partly-matching strings are categorized as two types: common strings and unique strings. An indicative example of the common strings is "school" that is commonly used in the education domain. These common strings do not provide any unique meaning of applicable terms. The threshold is set as three so that strings occurring more than three times at the LR level in one of the ontologies are considered as no matching strings. Terms containing uniquely matched strings which occur less than four times in one of the ontologies are enlisted as partly-matching terms.

2) Graph-based matching of the partly-matching terms in hierarchies below the 100 % matched nodes

The partly-matching terms identified in the string-based matching go through the graph-based matching process. The graph-based matching is based on the nodes that are categorized as the 100% matching terms and goes through from nodes situated at the lower part of the hierarchies to the upper part. For example, "university" and "higher education" are identified as 100 % matching pairs from the strings-based matching and the "university" is situated under the "higher education". Hence the graph-based matching first tries to identify the partly-matching terms listed under "university" nodes in the ontologies. In this example, "**master** degree course (JP)" and "Candidatus and **master** program (DK)" as well as "**doctor** degree course" and "**doctor** of philosophy program" are registered under the "university" nodes in the two ontologies. Hence, these are considered as corresponding pairs.

3) The internal structure (FSs) mapping of terms in the hierarchies below the matched nodes

The next step is to map the internal structure (set of feature specifications) of terms in the hierarchies below the already matched nodes. For example, at the hierarchies under the nodes "higher education", e.g. "specialized training college, specialist degree (JP)" and "academy profession degree program (DK)" are left behind. Presuming that these FSs play a similar role as DCFs described in Chapter 3, terms in the two ontologies (representing the SL- and TL concept system) having as many common FSs as possible can be assumed to be equivalent. Table 1 shows that e.g. the Japanese "specialist degree" is targeted for "high school graduates" whose age is "above 16 years". It means that the Japanese FS2 and FS4 can logically match with the Danish FS1 "19 years old or above". All other FSs besides the Japanese FS3 correspond with each other. Hence these two terms can be considered as corresponding translations.

Table 1: FSs mapping results

LR	FS 1	FS2	FS3	FS4	FS5	FS6
Specialist degree (JP)	PATH: traditional academic	**GENERA-TION: 16 years old or above**	OBLIGATION : non-compulsory	**ELIGIBILIT Y: high school graduate**	PURPOSE: practical	LENGTH OF EDUCATION: 2-3 years
Academy profession degree program (DK)	**GENERA-TION: 19 years old or above**	CATEGORY: academic study	LENGTH OF EDUCATION: 2 years	PURPOSE: practical		

The internal structure mapping can be repeated for terms in the hierarchies below the newly matched nodes. This process is repeated from the lower nodes to the top nodes of the hierarchies.

4.2 Results and evaluation

In this experiment, in total 42 terms and 49 terms are registered on the Japanese- and Danish education system ontologies, respectively. Among these, 14 corresponding English expressions that are mapped as equivalent based on the terminological ontologies are enlisted (Table 2). Since each English equivalent term carries its original expression (or its synonym) either in Japanese or Danish, original Japanese terms equivalent to Danish terms are considered as translation candidates.

Table 2: Potential translation list

Original Japanese	English (JP)	English (DK)	Original Danish
義務教育 (小学校・中学校) *Gimu-kyoiku (sho-gakko – chu-gakko)*	Compulsory education (elementary and junior high school)	1-9 years under primary and lower secondary education	Grundskole
高等学校普通科 *Koto-Gakko futu-ka*	High school General course	General upper secondary education	Det almene gymnasium/ studentereksamen
高等学校工業科 *Koto-gakko kogyo-ka*	High school technical course	Higher technical exam	Højere teknisk eksamen
高等学校商業科 *Koto-gakko syogho-ka*	High school business course	Higher commercial exam	Højere handelseksamen
高等教育 *Koto-kyoiku*	Higher education	Higher education	Videregående uddannselse
大学 *daigaku*	University	University	Universitet
学部 *Gakubu*	Undergraduate department	University bachelor program	Bachelor uddannelse
大学院前期課程 *Daigakuin zenki-katei*	Graduate school master degree course	University candidates and master program	Kandidatuddannelse
大学院後期課程 *Daigakuin koki-katei*	Graduate school doctor degree course	University doctor of philosophy program	Ph.d uddannelse
専門学校 (専修学校専門課程) 専門士 *Senmon-gakko (Sensyu-gakko senmon-katei) Senmon-shi*	Specialized training collage specialist degree	Academy profession degree program	Erhvervsakademiuddannelse / synonym: kort videregående uddannselse
専門学校 (専修学校専門課程) 高度専門士 *Senmon-gakko (sensyu-gakko senmon-katei) Kodo-senmon-shi*	Specialized training collage high level specialist degree	Professional bachelor program	Profession bacheloruddannelse/ synonym: mellemlang videregående uddannselse
幼稚園 *Yochien*	Kindergarten	Kindergarten	Børnehave
高等専修学校 *Koto-sensyu-gakko*	Specialized training school upper secondary course	Vocational education and training	Erhvervsuddannelse

Table 3: Translator's assessment

Original Japanese concept	Danish expression 1	Danish expression 2	Answer from translator
義務教育 (小学校・中学校)	Undervisningspligten (grundskole – junior høj)	**Grundskole**	Obligatorisk skolegang (grundskole)
高等学校普通科	Studentereksamen, infantry	**Det almene gymnasium/ synonym: studentereksamen**	**Det almene gymnasium**
高等学校工業科	**Højere teknisk eksamen**	Studentereksamen, Institute for industry	Teknisk gymnasium
高等学校商業科	Studentereksamen, Department of commerce	**Højere handelseksamen**	Handelsgymnasium
高等教育	**Videregående uddannselse**	Højere uddannelse	**Videregående uddannelse**
大学	College	**Universitet**	**Universitet**
学部	Undergraduate	**Bachelor uddannelse**	**Bacheloruddannelse**
大学院前期課程	**Kandidat uddannelse**	Graduate Master´s program	**Kandidatuddannelse**
大学院後期課程	Ph.d kursus	**Ph.d uddannelse**	**Ph.d.-uddannelse**
専門学校 (専修学校専門課程) 専門士コース	**Erhvervsakademiuddannelse / synonym: kort videregående uddannselse**	Erhvervsskole (faglig skole specialiserede kurser) certified professional	**Kort videregående uddannelse**
専門学校 (専修学校専門課程) 高度専門士コース	Erhvervsskole (faglig skole specialiserede kurser) certified advanced professional	**Profession bacheloruddannelse/ synonym: mellemlang videregående uddannselse**	**Mellemlang videregående uddannelse**
高等専修学校	**Erhvervsuddannelse**	Faglig high school	**Erhvervsuddannelse**

For assessing these results, a bilingual Dane who holds a Master's degree in the Japanese language, and has experiences in teaching Japanese and has been living in Japan for more than 15 years is appointed as evaluator. The evaluator is asked to perform the following task: *Select*

the most suitable Danish expression for an original Japanese concept from two choices; when no suitable expression is identified, propose a suitable translation for the Japanese concept. For alternative choices, the Google machine translations [4] (MT) of the Japanese original expressions are applied. The evaluation result is shown in Table 3. The bold letters indicate the results obtained from the ontology mapping. In addition, all the alternative translations, proposed by the evaluator, which contain some strings that are part of the translations obtained from the ontology mapping are underlined in Table 3. Among twelve Japanese terms, nine translations obtained from the ontology mapping are selected as suitable translations. All of the alternative Danish translations proposed for the remaining three concepts by the evaluator contain some strings obtained from the ontology mapping, while none of the Google translations are selected.

5 Discussion

5.1 Mapping Algorithms

In this experiment, the ontologies have semi-automatically been developed from the text corpora based on the principles of the terminological ontology. One critical point is the selection of subdividing dimensions and their values: How to select subdividing dimensions and their values that are commonly used for mapping the two ontologies. The more common names and characteristics of dimensions in the two ontologies are, the easier the mapping becomes. If considering the full automatic ontology development from text corpora, these issues should be carefully considered.

The ontology mapping has been manually performed in this experiment. A major challenge will be to identify a technique to automatically map FSs. At the current stage, the FS mapping involves human decisions in many ways. First of all, contexts of the subdividing dimensions and their values are inherently understood by humans. In order to automate the FS mapping, it at least has to employ the morphological matching technique for identifying corresponding strings in subdividing dimensions and their values. This means that the aforementioned problem concerning the selection of subdividing dimensions and values again becomes a critical issue. In addition, some complex situations involving human decision making have been identified in this experiment. For example, GENERATION “16 or above” implies “19 or above”. However, this cannot lead to the conclusion that these two values are corresponding. In the manual mapping, ELIGIBILITY “high-school graduate” can be an important clue to provide the fact of “3 years of education” from “16 years old”, which equals to “19 years or above”. In order to achieve semi-automatic or automatic matching algorithms, these kinds of issues should be dealt with in addition to the logical formalization of the procedures performed in this experiment. An obvious next step could be to implement semi-automatic FS mapping algorithms based on this procedure by employing already existing algorithms such as the Formal Concept Analysis (FCA) known as the concept lattice (Ganter and Wille, 1999).

5.2 Evaluation Methods

The evaluation method employed in this experiment involved only one evaluator assessing the translations obtained from the Google MT and from the ontology mapping. The first question is whether the comparison with the Google MT provides any meaningful results. Computational linguists and some psycholinguists may argue that frequency of word usage influences the translation performance. From this point of view, the results indicate that, for language combinations that have limited linguistic resources, the frequency approach, that is to say the statistical machine translation approach, may possess a disadvantage. Another critical issue is that only one evaluator was appointed for the evaluation. This may obviously not be sufficient

[4] http://translate.google.com/# as of June 8th, 2010

to convince about the validity of the results. The most important point of this project is whether the terminological ontology mapping conveys the original conceptual meanings of SL terms into TL translations compared with other translation approaches. This has not yet been proven by this single evaluation. Some psycholinguistic approaches, e.g. testing TL translation readers (for assessing whether a proposed translation evokes the original conceptual meaning), should be considered in future research.

In addition, objective and systematized evaluation methods should be developed in order to assess semi-automatic or automatic matching algorithms that are to be developed in the near future. One solution could be to participate in the Ontology Alignment Evaluation Initiative[5]. However, the characteristics of the ontology matching in this project are rather application specific. Hence, this issue should be investigated further.

6 Conclusions

In this work, the principles of terminological ontology development and procedures for mapping two domain-specific terminological ontologies are proposed. The results obtained from the ontology mapping are compared with translations obtained from the Google MT and with human translations. Among twelve Japanese terms, nine translations obtained from the ontology mapping are selected as suitable translations. All of the alternative Danish translations proposed for the remaining three concepts by the evaluator contain some strings obtained from the ontology mapping, while none of the Google translations are selected.

References

Carpenter B., 1992. *The Logic of Typed Feature Structures.* Cambridge University Press

De Groot, A. M. B. 1997. *The cognitive study of translation and interpretation: Three approaches.* In J.H. Danks, G. M Shreve, S. B. Fountain, & M. K. McBeath (Eds), Cognitive processes in translation and interpretation, Thousand Oaks, CA: Sage Publications, 25-56.

Euzenat J. and Shvaiko P. 2007. *Ontology Matching.* Springer-Verlag Berlin Heidelberg, 73-116

ISO 704:2000. Terminology work – Principles and methods. International Standards Organisation.

Ganter B and Wille R., 1999. *Formal Concept Analysis* Mathmatical Foundatins, Springer

Kroll, J. F. & Stewart, E. 1994. *Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations.* Journal of Memory and Language, 33, 149-174.

Madsen, B.N., Thomsen, H.E. and Vikner, C. 2004a. *Principles of a system for terminological concept modelling.* In Proceedings of the 4th International Conference on Language Resources and Evaluation. ELRA, 15-19.

Madsen, B.N., Thomsen, H.E. and Vikner, C. 2004b. Comparison of principles applying to domain specific versus general ontologies. In *Proceedngs of Ontologies and Lexical Ressources in Distributed Environments 2004*. ELRA, 90-95.

Madsen, B.N., Thomsen, H.E. and Vikner, C. 2005. Multidimensionality in terminological concept modelling. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, Copenhagen, 161-173.

Potter, M.C., So, K.F., von Eckardt, B. and Feldman, L.B. 1984. *Lexical and conceptual representation in beginning and proficient bilinguals.* Journal of Verbal Learning and Verbal Behavior, 23, 23-38.

Sager J.C., 1990. *A Practical Course in Terminology Processing.* Amsterdam: John Benjamins. 55-97.

[5] http://oaei.ontologymatching.org