

Alignment of Remote Cultures

In contrast to the Relevance Theory of Communication

Fumiko Kano Glückstad

Dept. of International Language Studies and Computational Linguistics
Copenhagen Business School, CBS
Frederiksberg, Denmark
e-mail: fkg.isv@cbs.dk

Abstract— As the role of ontology in a multilingual setting becomes important to Semantic Web development, it becomes necessary to understand and model how an original conceptual meaning of a Source Language word is conveyed into a Target Language translation. Terminological ontology [1] is a tool used for knowledge sharing and domain-specific translation, and could potentially be suitable for simulating the cognitive models explaining real-world inter-cultural communication scenarios. In this paper, a framework referred to as the Relevance Theory of Communication [2] is contrasted to an empirical study applying Tversky's contrast model [3] to data-sets obtained from the terminological ontology. The results indicate that the alignment of two language-dependent terminological ontologies is a potential method for optimizing the relevance required in inter-cultural communication, in other words, for identifying corresponding concepts existing in two remote cultures.

Keywords - Terminology, Ontology, Inter-cultural communication, Set-theory, Similarity, Relevance, Translation

I. INTRODUCTION

The role of ontology in a multilingual setting is an emerging challenge for Semantic Web development. As a consequence, there are several major ongoing projects, such as MONNET on Multilingual Ontologies for Networked Knowledge project [4] and KYOTO on Knowledge-Yielding Ontologies for Transition-Based Organization [5]. Though both projects deal with translation of terms from a Source Language (SL) to a Target Language (TL), they focus on linking lexical data through an interoperable common ontology rather than on optimizing relevance between concepts that are potentially measurable based on diverse models derived from the cognitive theory.

Another major project, the Language Grid [6] is an infrastructure that combines language resources such as bilingual dictionaries and language processing tools. Its technology is based on an ontology called Language Service Ontology and Semantic Web Service Technology. These two technologies enable one to combine bilingual dictionaries, e.g. Japanese-English and English-Danish dictionaries, and natural language processing tools from all over the world. Thus translations between languages, e.g. Japanese-Danish, can be realized through the transitive translation using

English as a pivot language. However, this approach focuses on combining the existing language resources rather than the cognitive process required in human translation. A question is how well such a pivot translation can convey an original conceptual meaning of an SL word into a TL translation.

Terminological Ontology (TO) is a domain-specific ontology used for knowledge sharing [1], which normally is applied in terminology work, cf. for example [7]. The unique points of TO that differentiate it from other types of ontologies are feature specifications and subdivision criteria [8]. A feature specification consists of a feature dimension and its value. Thus, a representation of a whole concept is a feature structure, i.e. a set of feature specifications corresponding to the unique set of characteristics that constitutes that particular concept [1][8]. Terminological ontologists argue that concepts are defined in a language dependent context, and therefore TO is language dependent. TO is developed within a knowledge sharing community, then dynamically updated and validated. If it is necessary to share knowledge with other communities, TOs developed in different communities should be compared, aligned and merged as needed. While the aforementioned two mainstream projects, MONNET and KYOTO, both deal with complex ontologies involving huge data-sets, TO usually handles smaller amounts of concepts.

Consequently, a point that should be emphasized in this work is that TO could potentially be a suitable method to apply and for simulating cognitive models explaining a real-world inter-cultural communication scenario. Thus, in this paper, a framework referred to as the Relevance Theory of Communication [2] is contrasted to an empirical study applying a classical psychological model [3] to data-sets obtained from TO [1]. The eventual purpose of this study is to propose an approach for identifying potential translation candidates by optimizing relevance between concepts in two remote languages such as a set of European and Asian languages. The expectation is that such translation candidates should provide better cognitive understandings of SL concept-meanings among a TL audience. However, in this preliminary study, the scope is limited to align ontologies constructed from English texts describing the educational systems in two remote cultures.

The rest of the paper is structured as follows: the Relevance Theory of Communication is reviewed in the next

section by use of a real-world example. Section 3 describes a classical psychological model applied to this preliminary study. In Section 4, the empirical analysis is performed to assess the potential of applying the models based on [3] to the terminological ontologies. Section 5 discusses findings and future work followed by conclusions in Section 6.

II. THE RELEVANCE THEORY OF COMMUNICATION

Imagine a situation where a non-native English speaking European and an Asian are debating in English about the issue of academic degree systems in their respective countries. While a Danish might be explaining about the Doctor of Science degree (the highest achievable academic doctor degree in Denmark after obtaining a Doctor of Philosophy that is considered a lower doctorate degree), a Japanese might be having the highest possible academic degree in Japan in his mind which is a Doctor of Philosophy degree (also frequently referred to as Doctor of Science in Japan). This imagined conversation shows a typical scenario revealing a deep inherent misconception between the two communicating parties since each of them have their own conceptual - and correct - understanding of the highest obtainable academic degree in their respective cultures.

This example may further create problems for a translator who is going to translate academic titles into the language of the other party. When a translator translates the term for the Danish Doctor of Science Degree into Japanese, the first condition he/she has to fulfill is that his/her translation should convey the same meaning as the original Danish meaning. Gutt [9] explains that this requires *the receptors to familiarize themselves with the context envisaged for the original text*. Now the question is, when a Japanese receptor is not familiar with the Danish language and its academic culture, how should this particular Danish academic title be translated into Japanese?

The proposal [9] of applying the Relevance Theory of Communication [2] might be a key to address this issue. This theory focuses on how people share thoughts with one another and views communication as principally an inferential process. It means that the essential task of the communicator is to produce a *stimulus* from which the audience can infer what set of thoughts or assumptions the communicator intends to convey [9]. Hence the second condition the translator has to fulfill is that his/her translation should explicitly provide a set of assumptions that are adequately relevant to the audience. The set of assumptions the translator provides should therefore be a stimulus that avoids unnecessary inferential processing effort required for the audience. The issue here is how the translator should create such *stimulus* (that is translation) optimally relevant to the audience. Assuming that both Danish and Japanese have their respective conceptual structures of the academic system rooted in their own culture, translation candidates that have optimally relevant relationships identified from these two conceptual systems could avoid the gratuitous inferential processing effort on the audience's part.

The optimization of the relevance between two concepts could be well explained by the cognitive model, Tversky's contrast model [3]. Thus the next section reviews Tversky's

model and considers how his model could be used in the context of optimizing the relevance of communication.

III. TVERSKY'S CONTRAST MODEL

The concept of similarity has a long history within the area of cognitive science. Tversky's view of similarity [3] is distinguished from the traditional theoretical analysis (c.f. [10]) on two points: 1) while the theoretical analysis of similarity relations has been dominated by the continuous metric space models, [3] argues that *the assessment of similarity between objects may be better described as a comparison of features rather than as the computation of metric distance between points*; and 2) although *similarity has been viewed by both philosophers and psychologists as a prime example of a symmetric relation*, the asymmetric similarity relation has been demonstrated in [3] based on several empirical evidences.

Based on these two points, [3] proposed a classic feature-set model of similarity as follows:

$$sim(a, b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A - B) + \beta \cdot f(B - A)} \quad (1)$$

Here, A and B are the feature sets of object *a* and object *b*. *f* denotes a measure over the feature sets. $(A \cap B)$ represents the sets of features present in both A and B, $(A - B)$ represents the sets of features present in A but not in B and $(B - A)$ represents the sets of features present in B but not in A. α and β are free parameters representing an asymmetric relationship between A and B. Since the similarity score in this equation is normalized, the obtained score lies between 0 and 1.

An interesting point is that the application of Tversky's model requires a *limited list of relevant features and the representation of an object as a collection of features that is viewed as a product of a prior process of extraction and compilation* [3]. In fact, the principle of TO in a way follows rigid rules of categorization. This can systematically extract the collection of features based on subdividing dimensions. Therefore, the hypothesis is that Tversky's model could be applicable to data-sets extracted from the terminological ontologies. Another important point in the context of the Relevance Theory of Communication is that translation should provide the set of assumptions that are adequately relevant to the audience. Considering that *similarity serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations* [3], the most similar concept to an SL concept, which is identified in the audience's culture through the feature matching, could be the set of assumptions which are adequately relevant to the audience. Thus, the second hypothesis is that the optimization of the relevance required in an inter-cultural communication can be achieved by aligning the feature specifications which constitute concepts in the two language-dependent terminological ontologies. In order to assess the first hypothesis, terminological ontologies are developed from corpora describing real-world concepts in the two remote

cultures. The similarity score of the selected concepts are computed by applying Tversky's model [3] based on the collection of features extracted from these ontologies. This is dealt with in the next section.

IV. ONTOLOGY ALIGNMENT

A. Corpora

Texts describing the Japanese educational system have been identified from the "Multilingual Living Information" site provided by the Council of Local Authorities for International Relations and from a pamphlet entitled "Higher Education in Japan" published by the Japanese Ministry of Education, Culture, Sports, Science and Technology. For the Danish educational system, documents that are downloaded from the Eurydice web-site published by the Education, Audiovisual and Culture Executive Agency under the EU commission have been used as text corpus. All these documents are officially published in English by reliable authorities of each country. Thus, all English translated terms and expressions in their original languages are considered as official terms. It means that it is feasible to identify terminological expressions in an original language from documents published by the respective authorities. This enables one to eventually identify translation equivalences linking between, in this case, Danish and Japanese. In this study, only the English documents describing about language-dependent concepts in the two cultures are used as text corpora.

B. Ontology construction

The terms and their definitions describing the educational systems in each country are manually identified from the respective English corpora. Based on these terms and their definitions, terminological ontologies representing the educational system in each of the two countries are developed using the Computer Aided Ontology Structuring prototype (CAOS) that is based on TO principles defined in [1]. As described in Section 1, the uniqueness of TO is feature specifications and subdivision criteria [8]. A feature specification is presented as attribute-value pair. Thus, a representation of a whole concept is a feature structure, i.e. a set of feature specifications corresponding to the unique set of characteristics that constitutes that particular concept [1].

The use of feature specifications is subject to principles and constraints described in detail in [1]. Most importantly, a concept automatically inherits all feature specifications of its superordinate concepts. Secondly, polyhierarchy is allowed so that one concept may be related to two or more superordinate concepts. On the other hand, subdivision criteria that have been used for many years in terminology work are strictly implemented in TO by introducing dimensions and dimension specifications [1][8]. This enables the CAOS prototype to perform consistency checking which helps in constructing ontologies [1]. A dimension of a concept is an attribute occurring in a non-inherited feature specification of one or more of its subordinate concepts [1][8]. Values of the dimension allow a distinction among sub-concepts of the concept in question. For example, a

dimension of a concept "higher education" is [TITLE] whose values are [associate | bachelor]. These dimension values distinguish the sub-concepts: "junior college" and "university (undergraduate)". The dimension can only occur on sister concepts and a given value can only appear on one of these sister concepts. In this way 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 feature specification [1][8].

By using the CAOS prototype that performs the consistency checking of TO principles, the two educational ontologies are developed based on the terms and definitions manually extracted from the corpora.

C. Application of Tversky's contrast model

Based on the TO principles, the Japanese and Danish educational ontologies respectively containing 40 and 54 concepts are developed. All concepts consist of features systematically extracted from the terminological ontology. For example, the Japanese term, "specialized training college – specialist course", consists of features {formal education, require upper secondary exam, profession oriented education, 2-3 years program duration}. On the contrary, the Danish term, "short cycle education", consists of features {require upper secondary exam, project & research based, profession oriented education, administered by Ministry of Education, 1-3 years program duration, vocational college, earn credit to medium or longer cycle education}. These features existing in a Danish concept and a Japanese concept are used for computing a similarity score based on Tversky's model (equation 1). In order to apply this model, synonymous expressions of features extracted from the country specific corpora are approximately standardized by hand. Accordingly, in case of the aforementioned examples, $(A \cap B)$ is 2, since the features {require upper secondary exam, profession oriented education} exist in both the Danish and Japanese terms. As for the asymmetric similarity, it is defined in [3] that *if $sim(a,b)$ is interpreted as the degree to which a is similar to b , then a is the subject of the comparison and b is the referent. Hence the features of the subject are weighted more heavily than the features of the referent.* When considering a translation scenario, translators' task is to identify a concept in audiences' conceptual structure that is optimally relevant to the concept in the SL. It means that the stimulus selected by a translator should to the maximum extent be similar to a concept in the SL concept. Therefore, the features of a stimulus should be weighted more heavily than the ones of an SL concept in accordance to [3]. Hence, the asymmetric parameters are manually set as $\alpha=0.7$ and $\beta=0.3$ in this empirical study. It means that the similarity score becomes smaller, when $(A-B)$ is larger than $(B-A)$ in the equation (1). In case of the aforementioned example of "specialized training college – specialist course" and "short cycle education", $(A-B)$ is 5 and $(B-A)$ is 2 when the Danish term is set as variant. On the other hand, $(A-B)$ is 2 and $(B-A)$ is 5, when the Japanese term is set as variant. Accordingly, the asymmetric similarity scores are, 0.328 and 0.408, respectively. In this way, the asymmetric similarity scores are computed in all possible

combinations between the Japanese and the Danish concepts. Examples of the mapped concepts based on the similarity scores are illustrated in Figure 1 and 2. The concepts listed in Figure 1 and 2 are automatically selected based on the similarity scores. It means that Tversky's model roughly categorizes clusters that are more or less similar to each other. In these figures, the bold lines are the highest scores equal to or over 0.5 obtained for each SL concept, and the slim lines are the highest score lower than 0.5, or the second highest scores equal or higher than 0.5. The bold dotted lines are ideal corresponding concepts and the slim dotted lines are acceptable corresponding concepts defined by human.

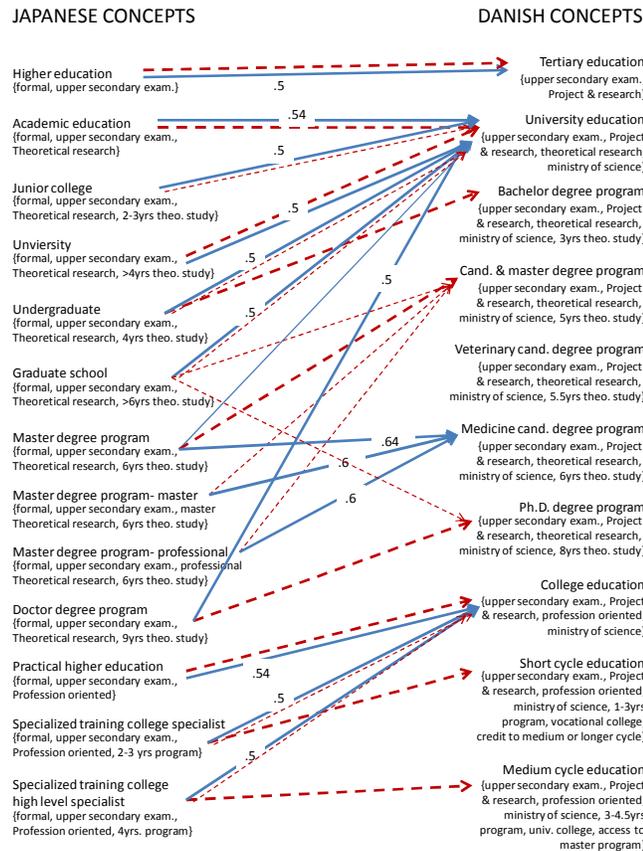


Figure 1. Tversky's similarity score (Danish as variant)

Based on the aforementioned criteria, precision and recall scores are computed based on the following equations.

$$\text{precision} = \frac{\{\text{corresponding pairs}\} \cap \{\text{mapped pairs}\}}{\{\text{mapped pairs}\}} \quad (2)$$

$$\text{recall} = \frac{\{\text{corresponding pairs}\} \cap \{\text{mapped pairs}\}}{\{\text{corresponding pairs}\}} \quad (3)$$

The precision and recall scores are calculated in two ways. For the first scores (referred to as "strict"), only the ideal

corresponding pairs (bold dotted lines) are considered as corresponding concepts and only the highest scores (bold lines) are considered as most similar concepts. For the second scores (referred to as "relaxed"), the acceptable corresponding pairs (slim dotted lines) are included in addition to the ideal corresponding pairs and the highest score lower than 0.5, or the second highest scores equal or higher than 0.5 (slim lines) are included in addition to the most similar concepts. The results are summarized in Table 1.

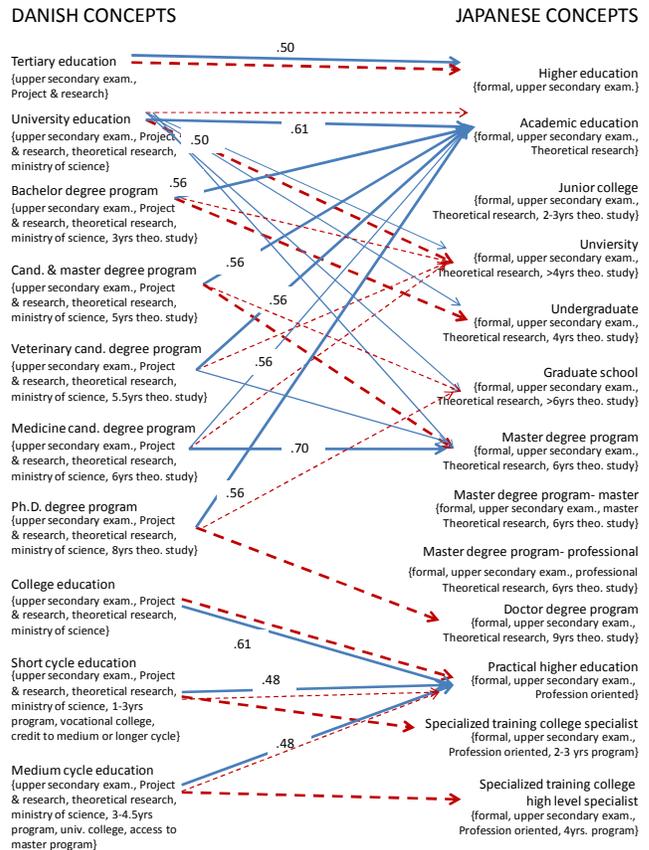


Figure 2. Tversky's similarity score (Japanese as variant)

TABLE I. PRECISION AND RECALL

	Strict alignment				Relaxed alignment			
	Corresponding pairs	Mapped pairs	Precision	Recall	Corresponding pairs	Mapped pairs	Precision	Recall
DA as variant	26	36	27.8 %	38.5 %	72	77	44.2 %	47.2 %
	10				34			
JA as variant	26	32	28.1 %	34.6 %	53	62	25.8 %	30.2 %
	9				16			

V. DISCUSSION

A. Findings

In this empirical study, the terminological ontologies are constructed from English texts describing the educational systems in two remote cultures. As Figure 1 and 2 illustrate, the English expressions of concepts in the two ontologies are in most cases not identical. This indicates that the English educational terminology used in respective knowledge sharing communities is immensely dissimilar, even though the educational concepts existing in the two countries are relatively similar. From this fact, it can be elaborated that it is a very complicated task to link concepts in two remote languages. This is because language resources having direct links between two remote languages are usually very limited, and therefore it often requires a pivot translation using English both for dictionary-based human translations and for statistically-based machine translations. This also emphasizes the necessity for carefully analyzing how a concept in one culture can be conveyed to a person in another culture through English as *lingua franca*.

In this study, the feature-based asymmetric similarity has been computed for all combinations of concepts in two cultures in both directions. Accordingly, concepts in the two ontologies were approximately clustered into 5 groups. Figure 1 and 2 show one of the clusters automatically extracted from this process. These figures indicate that concepts referring to education for people who have completed the secondary education are adequately extracted from the two ontologies. This can be explained in the following way: The principles of Terminological Ontology - e.g. inheritance of feature specifications, uniqueness of dimensions, and uniqueness of primary feature specifications - enable features to reflect the structure of ontologies. For example, a feature value, "upper secondary exam" in Figure 1 and 2 has been specified only once when defining the concept "higher education" and "tertiary education" respectively in the Japanese and the Danish ontologies. Since all of the sub-concepts under these two terms inherit this feature value, the concepts referring to education for people who completed the upper secondary exam have systematically been extracted and grouped as a cluster.

The computation of the asymmetric similarity scores for all combinations of concepts within each cluster resulted in identifying a concept in a target ontology that has the highest optimal relevance to a source concept. Figure 1 and 2 show that, in most cases, the identified optimal stimuli based on the asymmetric score was the most general term located at the highest node within the cluster. According to the principle of TO, when a concept is subdivided into several sub-concepts based on a dimension, an extra feature is added to each sub-concept. Hence it is always the case that concepts having more features are more specific sub-concepts. It means that the lower a concept is located in the ontology, the more features the concept inherits from superordinate concepts. If dimensions and their values at the lower part of the two ontologies are not consistent, all the inherited features are simply noise in the data-sets. Therefore, the identified optimal stimuli were in most cases

more general and abstract terms rather than specific terms. The positive interpretation of this phenomenon could be that Tversky's model is applicable to identify corresponding pairs with less noise, in other words, pairs that optimally share common features with less noise. The negative interpretation could be that Tversky's model has limitations in identifying corresponding pairs at the optimally specific level. Considering communication in the real world, it is not incorrect to say that the relevance required in the communication is achieved in this way, since people can often achieve mutual understandings much easier at a reasonably general level than at a very specific level. However, there is a room for further investigating this phenomenon in the future.

Another notable point is that, the asymmetric similarity measure based on common- and distinctive features seems to be a very useful approach for the translation and intercultural communication. The results shown in Table 1 indicate that the precision and recall scores for Danish as variant are generally higher than Japanese as variant. The reason could be that the Japanese ontology has extra layers of the hierarchy compared to the Danish one. This made it difficult to identify a Japanese concept as stimulus. Even though concepts in two cultures are mapped to each other, it is not necessarily true that a translational equivalence holds both ways, if the two concepts are not 100% identical. Furthermore, in culturally dependent domains, some concepts simply do not exist in another culture. In such situations, reasonably abstract concepts identified as a corresponding pair in the target ontology could be considered as the optimal translation. This cannot be achieved by symmetric similarity measures.

Finally, a major problem has also been identified in this empirical study. In a cluster consisting of concepts referring to the Danish continuing education and Japanese alternative education, some mismatches have been identified. This is because the Japanese education system has been classified into two categories, "formal education" and "alternative education" at the highest level of the ontological hierarchy, while the Danish education system has been classified based on the "age and entrance qualification" dimension at the highest level. Under the Danish continuing education concept, there are two types of education, "formal education" and "non-formal education". Since the feature value "formal education" in the two ontologies was considered as a common feature, all the Japanese formal educations that are from the primary education to the higher education and the Danish formal educations under the Danish continuing education targeted for adults were considered as corresponding concepts. This strongly influenced the precision and recall scores in Table 1. This example indicates that the selection of feature dimensions and values upon the development of ontology is a very sensitive issue.

B. Future challenges

One of the challenges pointed out from the aforementioned findings is to identify reasonably specific corresponding concepts from noisy data-sets. One approach is to investigate each situation where features are considered

as either noisy or useful data. It is pointed out in [11] that Tversky's contrast model allows any kind of features and any feature weights whatsoever. However, to improve the mapping quality, some semantic questions such as what qualifies as a feature and what determines the feature weights should be investigated. Considering the aforementioned example, even though the feature value "formal education" is considered as a common feature, if other features in the Danish and Japanese concepts consist of features that hold a conflicting relationship, e.g. one has a feature "6 years old or above" and the other "adult", the impact of common feature should be smaller. How to weight each feature may be a key to improving the mapping quality. The Bayesian model proposed in [11] explains some aspects of the origins and the dynamics of feature weights. Another question raised in [11] is that Tversky's formal theory does not explain why and how the two subsets of distinctive features (A-B) and (B-A) should be given different weights. In fact, the parameters used in this empirical study have been randomly selected. This issue is dealt with in the Bayesian model as well. Hence, one way to approach this is to extend Tversky's model based on the idea proposed in [11] and apply it to the same data-sets.

Another way of approaching the challenge is to investigate how feature dimensions and values are selected during the ontology construction phase. Considering the technical aspects of the ontology matching, handling of the ontology construction may not be the scope for the improvement of matching results. However, for the purpose of translation and inter-cultural communication of domain specific terms that are constrained by a cultural boundary (such as education system, social system, legal system, financial system), the standardization is considered as an important process for sharing knowledge across country borderlines. In fact, the documents describing the educational system for the majority of EU member countries are based on the standardized classification and terms on the Eurydice web-site, since the EU commission has employed English terminology that is standardized based on the International Standard Classification of Education (ISCED) defined by the United Nations Educational, Scientific and Cultural Organization. Therefore, it may be much easier to align the educational system ontologies within the EU member countries. Further investigation and comparison of this empirical study with data-sets obtained from standardization practices as well as from the translation practices in the real world would contribute to the optimization of the relevance required in an inter-cultural communication scenario.

VI. CONCLUSIONS

In this paper, a framework referred to as the Relevance Theory of Communication proposed by [2] is contrasted to an empirical study applying Tversky's contrast model [3] to

data-sets obtained from the Terminological Ontology method [1]. The results demonstrate that the application of Tversky's contrast model to the data-sets extracted from the terminological ontologies is to a certain degree an effective approach. However, further investigations are needed to assess whether the optimization of the relevance required in an inter-cultural communication can be achieved by aligning the feature specifications which constitute concepts in the two language-dependent terminological ontologies.

ACKNOWLEDGMENT

I would like to express my gratitude to my supervisors, Hanne Erdman Thomsen and Daniel Hardt for their valuable guidance and support on my Ph.D. project.

REFERENCES

- [1] B.N. Madsen, H.E. Thomsen, and C. Vikner, "Principles of a system for terminological concept modelling," Proc. The 4th International Conference on Language Resources and Evaluation, ELRA, 2004, pp. 15-19
- [2] D. Sperber and D. Wilson, "Relevance: Communication and Cognition", Blackwell, Oxford, 1986
- [3] A., Tversky, "Features of similarity," Psychological Review 84, 1977, pp. 327-352
- [4] T. Declerck, H.U. Krieger, S.M. Thomas, P. Buitelaar, S. O'Riain, T. Wunner, G. Maguet, J. McCrae, D. Spohr, and E. Montiel-Ponsoda, "Ontology-based multilingual access to financial reports for sharing business knowledge across Europe," Internal Financial Control Assessment Applying Multilingual Ontology Framework, J. Rooz, J. Ivanyos, Eds. Budapest: HVG Press, 2010, pp. 67-76
- [5] P. Vossen, E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, and M. Tescon, "KYOTO: A system for mining, structuring and distributing knowledge across languages and cultures," Proc. The 6th International Conference on Language Resources and Evaluation, Morocco, 2008, pp. 1462-1469
- [6] T. Ishida, "Language Grid: An infrastructure for intercultural collaboration," Proc. IEEE/IPSJ Symposium on Applications and the Internet (SAINT 06), 2006, pp. 96-100
- [7] ISO 704, "Terminology work — Principles and methods," Genève: ISO, 2000
- [8] B.N. Madsen, H.E. Thomsen, C. Vikner, "Comparison of principles applying to domain specific versus general ontologies," OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, A. Oltramari, P. Paggio, A. Gangemi, M.T. Pazienza, N. Calzolari, B.S. Pedersen, K. Simov, Eds. ELRA, 2004, pp. 90-95
- [9] E.A. Guut, "A theoretical account of translation – without a translation theory," Target: International Journal of Translation Studies, 2/2, 1990, pp.135-164
- [10] R.N. Shepard, "Toward a universal law of generalization for psychological science," Science, New Series, Vol. 237, No. 4820, 1987, pp. 1317-1323
- [11] J.B. Tenenbaum, and T.L. Griffiths, "Generalization, similarity, and bayesian inference," Behavioral and Brain Sciences 24, 2001, pp. 629-640