Distributional and translational solutions to the visualization of semantic differences between translated and non-translated Dutch

Lore Vandevoorde, Pauline De Baets, Els Lefever, Koen Plevoets & Gert De Sutter
Research unit EQTIS and LT3, Department of Translation, Interpreting and Communication
Ghent University, Belgium
{Lore.Vandevoorde}; {Pauline.DeBaets}; {Els.Lefever}; {Koen.Plevoets}; {Gert.DeSutter}@UGent.be

***Abstract*—This paper aims to visualize the semantic field of inchoativity in Dutch, for both translated and non-translated language. Two methodological solutions, a context-based and a translation-based approach, will be assessed and consequently compared to each other. Such a comparison can possibly generate interesting insights into the accuracy of the results of both the context-based and the translation-based method.**

***Keywords*—Corpus-based Translation Studies, semantic mirroring, distributional semantics, translation-based, inchoativity**

I. INTRODUCTION

In Corpus-based Translation Studies, it is often assumed that translated texts incorporate typical linguistic features which distinguish them from non-translated, original texts [1]. The question whether these typical features of translated texts also exist on the semantic level has, however, rarely been raised. With this paper, we want to contribute to the under-researched field of semantics in translation by focusing on the visualization of a specific semantic field, viz. inchoativity, in both non-translated (original, source) language and translated language. More specifically, the aim of this paper is to assess two methodological solutions – a context-based approach versus a translation-based approach – for the visualization of semantic fields of non-translated and translated language.

II. BACKGROUND

In order to compare semantic fields across varieties – in our case, non-translated and translated language – we first have to be able to objectively generate semantic fields for each of the varieties. In lexical semantics and lexical variation studies (e.g. [2]), the idea that the meaning of a word can be deduced from the company it keeps [3, 4] has led to the advent of (semi-)automatic retrieval methods of semantically similar words such as latent semantic analysis [5], first and second order bag-of-words models [6] and the behavioral profiles method [7, 8]. These models are generally characterized as distributional, which means that they capture word meaning in relation to their context in large corpora. In word sense disambiguation, unsupervised corpus-based methods to disambiguate the different senses of a word are either based on the distributionalist hypothesis, or, alternatively, on the idea of translational equivalence [9] – the latter hypothesis is then based on the idea that a word can be known by the *translational* company it keeps. While distributional approaches are indeed widely applied in (theoretical) Lexical Semantics, methods that rely on translational equivalence as a meaning-structuring device have not yet had much uptake in this area of semantics. Admittedly, the distributional hypothesis has opened the way to a myriad of methodological possibilities and fine-grained analytical tools (which do not seem to have reached their limitations yet) so the 'need' to rely on an alternative hypothesis can seem somewhat obsolete. However, if one is interested in investigating the semantics of translated language (in comparison to non-translated language), the translational hypothesis might be an appropriate starting point. In addition, using an alternative method that aims at arriving at the same goal as a distributionalist analysis can generate interesting insights into the accuracy of the results of both the translational and the distributional method.

III. METHOD

A. *Translational method: SMM++*

In this paper, we thus compare a translational and a distributional method for semantic field visualization. A translational method has recently been proposed [10], drawing on the idea first uttered by Dyvik [11] that semantic relations can be deduced from overlapping sets of translations. A translation-based retrieval task, called Semantic Mirroring Method++ was developed on the basis on Dyvik's idea of semantic mirroring and was carried out as follows: by looking up the English (or French) translations of an initial lexeme ('beginnen' [to begin] – a central expression of inchoativity in Dutch) back and forth in a parallel corpus, 17 Dutch lexemes were selected on the basis of their semantic relatedness to the initial lexeme 'beginnen' ('aanvang' [commencement], 'begin' [beginning], 'beginnen' [to begin], 'eerst' [firstly], 'gaan' [to go], 'komen' [to come], 'krijgen' [to get], 'ontstaan' [to come into being], 'openen' [to open], 'oprichten' [to establish], 'opstarten'[to start up], 'opzetten' [to set up], 'start'[start], 'starten'[to start], 'van start gaan' [to take off], 'vanaf' [as from], 'worden'[to become]). Note that the application of this technique allows the researcher to select semantically related lexemes pertaining to different word classes. Three sets of data were subsequently extracted from the Dutch Parallel Corpus (see below): one set for non-translated language and two sets for translated language (one translated from English, a second translated from French), with each of the sets for translated language consisting of ± 1000 attestations (n=829 for Dutch translated from English, n = 1179 for Dutch translated from French) and > 2000 attestations for non-translated language (n = 2607 for original/non-translated Dutch). Every attestation in the data sets thus consisted of a sentence in Dutch containing one of the 17 Dutch lexemes (either as source or as target language lexeme) and a parallel sentence in English or French (either as the target or as the source sentence of the Dutch sentence). For each pair of parallel sentences, we annotated, in a separate file, the lemmatized Dutch lexeme together with its parallel lemmatized translation/source language lexeme. This two-column file was consequently transposed into a data matrix where the Dutch lexemes are represented in the rows and the translations/source language lexemes in the columns. The Dutch lexemes were subsequently visualized on the basis of their translational counterparts via the statistical technique of hierarchical cluster analysis [12, 13] which was carried out on the output of a correspondence analysis [14, 15]. The number of clusters was determined on the basis of pvclust [16], a technique that assesses the 'quality' of every node (i.e. each point where two branches join) in the cluster tree by calculating p-values (values between 0 and 1). We chose to apply an additional function of pvclust, i.e. pvrect, which cuts the tree at every highest node with a significant p-value. The visualized results (Figures 1, 3, 5) showed structural resemblances and small but noteworthy differences between the semantic fields of original texts and translations, as translations seem to 'flatten' meaning differences by which we mean that the central cluster containing 'beginnen' contains more lexemes in translated language (Figures 3 and 5) compared to non-translated language (Figure 1). We consider the cluster with 'beginnen' as the most central cluster, which might be interpreted as the prototypical center (this is confirmed by an additional analysis which calculates the average distances of each of the clusters to the zero-point of the semantic space). If more lexemes pertain to the prototypical center, this means that more lexemes are used to express the most prototypical sense of inchoativity. Had there been less lexemes in the prototypical center, this would have meant that only a few lexemes could be used to express the most prototypical sense of inchoativity with the other lexemes pertaining to other clusters expressing some kind of difference in meaning compared to the prototypical meaning. With more lexemes in the prototypical center, it is implied that all lexemes which pertain to the prototypical center are used in the more prototypical sense (lexemes within the prototypical center are more similar to each other than they are to lexemes in other clusters) and consequently 'lose' some of the meaning distinctions that were present in non-translated language. Although the use of translational data indeed seems to be a most useful tool to model semantic differences between language varieties in general and translated and original/source language in particular, the method does need further testing and comparison before additional statements can be made on the basis of this 'translational semantic approach'.

B. *Distributional method*

As a next step, we carry out a distributional analysis for the field of inchoativity and compare its outcome to the results of the translational approach to that same field. With this first comparative step, we hope to arrive at a better understanding of (i) the way in which the translation-based meaning structuration differs from the context-based one, and (ii) the extent to which the methods can be compared to each other. In order to allow for such a comparison, the data retrieval task for the distributional analysis is carried out in exactly the same way as was done for the translational approach, i.e. with the same initial lexeme 'beginnen' [to begin] and via the same retrieval method (Semantic Mirroring Method++) and applied to the same corpus, i.e., the Dutch Parallel Corpus – a ten million word, sentence aligned, both parallel and comparable corpus of Dutch, French and English [17]. This led to the creation of 3 sets of data, which are in itself identical to the data sets used for the translational approach. Whereas the translational method made use of the resulting cross-lingual, translational information present in these retrieved sets to create the data matrix, the distributional method uses monolingual, contextual information of the Dutch sentences (i.e. the immediate context words of the target lexeme). For the set of data representing non-translated language, the contextual information of the non-translated Dutch lexemes is used to create the data matrix; as for the set of data representing translated language, the target language contextual information of the data sets where Dutch is the target language of either French or English source language sentences is used for the creation of the data matrices. By means of LeTs Preprocess [19] , we automatically lemmatized all Dutch context sentences containing the 17 lexemes expressing inchoativity, and created frequency matrices of those 17 lexemes with their immediate context words. We retained only those context words that appeared more than once in the corpus in order to filter out idiosyncratic expressions. The distributional analysis was carried out on different context windows: 3 words to the left and to the right, 5 words to the left and the right and the sentence as a whole, following the bags-of-words model [6]. We did not exclude the top frequency terms - mostly function words for the time being; we do plan to carry out a distributional analysis that excludes those top frequency function words so as to compare the impact of the different parameters on the structure of the semantic field. The obtained matrices were then statistically analyzed in exactly the same way as was done for the translational approach to the same field of inchoativity: after a preliminary correspondence analysis allowing for dimension reduction and removal of noisy data [18], the resulting lexeme-coordinates were implemented into a hierarchical cluster analysis. It is important to note that thanks to the preliminary correspondence analysis, our clusters are not based on raw data since the correspondence analysis maps the data points on a reduced set of underlying dimensions and uses those dimensions as the basis for the subsequent hierarchical cluster analysis. Parallel to the translational visualization task, we chose to employ Euclidean distance and Ward's clustering algorithm. To ensure cluster stability, we let a bootstrap run over the data. Bootstrapping computes estimated standard errors by resampling the data set a specified number of times (3000 times in our case). It then calculates the specific statistics from each sample in order to arrive at the standard deviation of the sample.

IV. RESULTS

The results of the distributional analysis (Figures 2, 4 6) show that the general observations made on the basis of the translational method also hold on the basis of a distributional analysis: the semantic structure of translated (Figures 4 and 6) and non-translated language (Figure 2) for inchoativity show clear structural resemblances: for each visualization and regardless of the chosen method, the cluster containing 'beginnen' is the most central one in the analysis and represents the prototypical center – this is confirmed by an additional analysis which calculates the average distances of each of the clusters to the zero-point of the semantic space. Furthermore, and still regardless of the method, this central cluster also becomes larger in translated language (except for the distributional approach to Dutch translated from French), allowing thus for less meaning differentiation in the overall structure of translated language compared to non-translated language (although it needs to be noted that the distributional analysis of translated Dutch from French is remarkably similar to the distributional analysis of non-translated Dutch). When observing the members of each cluster in more detail, we can say that the differentiation between lexemes emphasizing the dynamic nature of the action (cluster around 'starten' [to start]) versus the lexemes emphasizing the state-after-onset of the action (cluster around 'beginnen' [to begin]) – a difference pointed out for English by Divjak & Gries [8] and which assumingly also exists for Dutch – is more clearly foregrounded by the distributional analysis. Although this distinction is also laid bare by the translational analysis for Dutch translated from English (Figure 5), the translational analyses seem to be better at detecting other types of distinctions in the semantic structure: the translational analysis often seems to point towards clusters of near-synonymous lexemes (e.g., Figure 1 contains clusters such as " 'oprichten' [to found] and 'opzetten' [to set up]", " 'ontstaan' [to come into being] and 'openen' [to open]", " 'aanvang' [commencement], 'begin' [beginning], 'start' [start]") whereas the distributional analysis lays bare in a more clear way a 'deeper' distinction such as the one between 'action' and 'state after onset' related lexemes (compare the cited clusters of Figure 1 to Figure 2, where none of the near-synonymous clusters are present as such but a cluster of all dynamic, action-like lexemes is formed instead (start [start], starten [to start], oprichten [to set up], openen [to open], opstarten [to start up], opzetten [to set up]).

V. CONCLUSION

Our conclusions can be formulated on two levels. First, the distributional analysis presented in this paper confirms the hypothesis that semantic fields representing translated language seem to present less meaning differentiation than the fields representing non-translated language (especially when translated from English), underpinning the previously uttered idea that meaning is somehow flattened in translation. Second, a first comparison of the translational with the distributional approach seems to indicate that the way in which each of the approaches captures and structures meaning is – unsurprisingly – different (which inevitably leads to the question to what point a comparison of the two methods can be maintained) – and leads to structurations of the field of inchoativity along different though meaningful lines: the translational approach leads to a clustering that favors near-synonymous groupings, whereas the distributional method seems to favor a clustering based on a 'deeper' linguistic distinction. Whether the researcher can benefit more from the one or the other analysis will depend on the type of research question he wishes to answer. Obviously, more research will still be needed to determine the stability of the observed differences between the methods.

Fig. 1. Translational approach to the field of inchoativity: non-translated Dutch

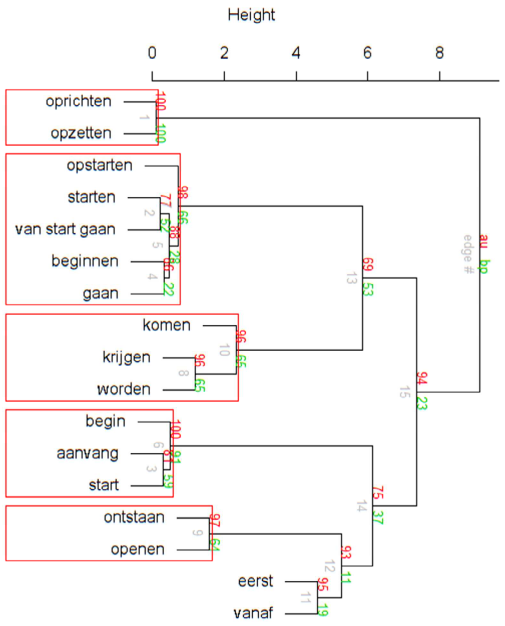


Fig. 2. Distributional approach to the field of inchoativity: non-translated Dutch

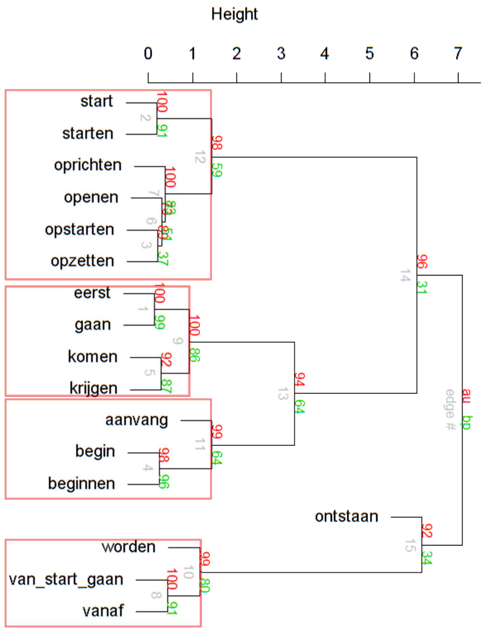


Fig. 3. Translational approach to the field of inchoativity: translated Dutch (French source language)

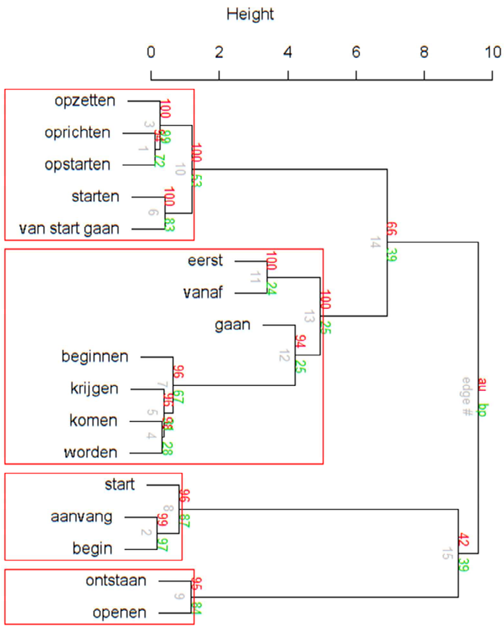


Fig. 4. Distributional approach to the field of inchoativity: translated Dutch (French source language)

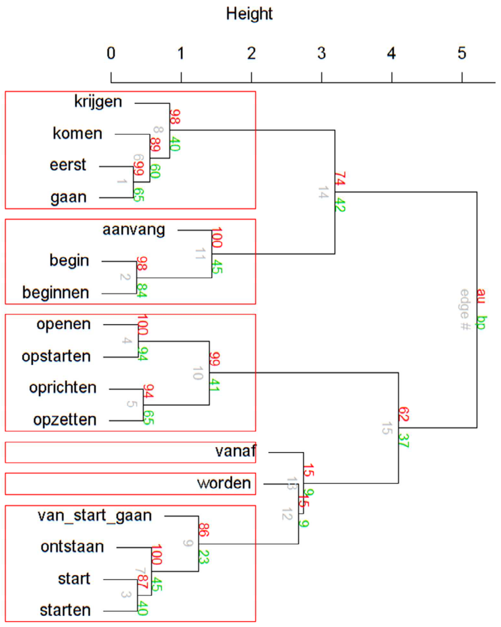


Fig. 5. Translational approach to the field of inchoativity: translated Dutch (English source language)

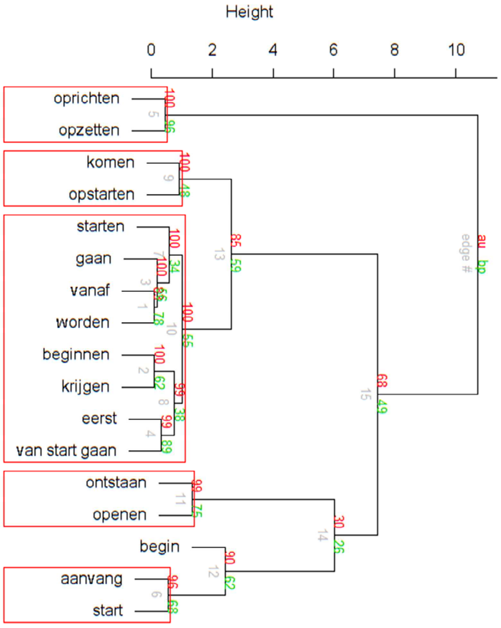
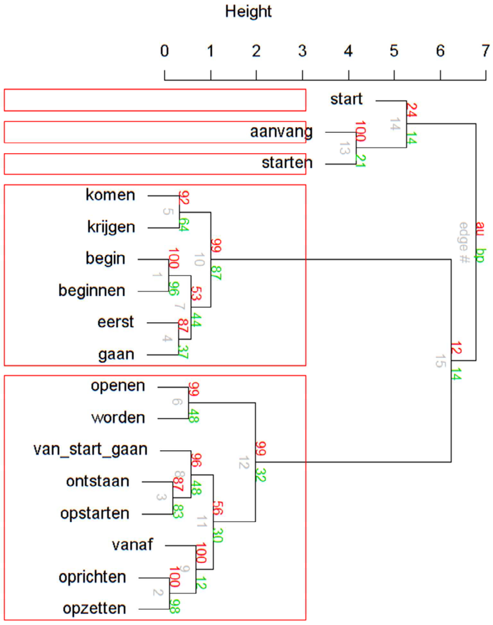


Fig. 6. Distributional approach to the field of inchoativity: translated Dutch (English source language)



References

[1] Baker, Mona. "Corpus Linguistics and Translation Studies. Implications and Applications.". In *Text and Technology. In Honour of John Sinclair.*, edited by Mona Baker, Gill Francis and Elena Tognini-Bonelli, 17-45. Philadelphia/Amsterdam: John Benjamins, 1993.

[2] Peirsman, Yves, Dirk Geeraerts, and Dirk Speelman. "The Automatic Identification of Lexical Variation between Language Varieties." *Journal of Natural Language Engineering* 16, no. 4 (2010): 469-91.

[3] Firth, John R. "A Synopsis of Linguistic Theory 1930-1955.". In *Studies in Linguistic Analysis*, edited by John R. Firth, 1-32. Oxford: Philological Society, 1957.

[4] Harris, Z. "Distributional structure". *Word* 10, no. 23 (1954): 146–162.

[5] Landauer, Thomas, and Susan Dumais. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge." *Psychological Review* 104, no. 2 (1997): 211-40.

[6] Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[7] Divjak, Dagmar, and Stefan Gries. "Ways of Trying in Russian. Clustering Behavioral Profiles." *Corpus Linguistics and Linguistic Theory* 2, no. 1 (2006): 23-60.

[8] Divjak, Dagmar, and Stefan Gries. "Corpus-Based Cognitive Semantics. A Contrastive Study of Phasal Verbs in English and Russian." In *Studies in Cognitive Corpus Linguistics*, edited by B. Lewandowska-Tomasczyk and Katarzyna Dziwirek, 273-96. Frankfurt am Main: Peter Lang, 2009.

[9] Agirre, Eneko, and Philip Edmonds. "Introduction." In *Word Sense Disambiguation.*, edited by Eneko Agirre and Philip Edmonds, 2007.

[10] Vandevoorde, Lore, Gert De Sutter, and Koen Plevoets. "On Semantic Differences between Translated and Non-Translated Dutch. Using Bidirectional Parallel Corpus Data for Measuring and Visualizing Distances between Lexemes in the Semantic Field of Inceptiveness." In *Empirical Translation Studies. Interdisciplinary Methodologies Explored*, edited by Ji Meng, 128-46. Sheffield & Bristol: Equinox, 2015.

[11] Dyvik, Helge. "A Translational Basis for Semantics." In *Corpora and Cross-Linguistic Research: Theory, Method, and Case Studies*, edited by S. Johansson and S. Oksefjell, 51-86. Amsterdam: Rodopi, 1998.

[12] Everitt, Brian S., Sabine Landau, Morven Leese, and Morven Stahl. *Cluster Analysis*. 5 ed.: Wiley, 2011.

[13] Divjak, Dagmar, and Nick Fieller. "Cluster Analysis. Finding Structure in Linguistic Data." In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy.*, edited by Dylan Glynn and Justyna Robinson, R., 405–41, 2014.

[14] Greenacre, Michael. *Correspondence Analysis in Practice, Second Edition.* . Boca Raton: Chapman & Hall/CRC, 2007.

[15] Lebart, Ludovic, André Salem, and Lisette Berry. *Exploring Textual Data*. Dordrecht: Kluwer Academic Publishers, 1998.

[16] Suzuki, Ryota, and Hidetoshi Shimodaira. "Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering." Bioinformatics 22, no. 12 (2006): 1540-42.

[17] Macken, Lieve, Orphée De Clercq, and Hans Paulussen. "Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus." *Meta* 56, no. 2 (2011).

[18] Ciampi, Antonio, Ana González Marcos, and Manuel Castejón Limas. "Correspondence Analysis and Two-Way Clustering." *SORT* 29, no. 1 (2005): 27-42.

[19] Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., & Hoste, V. "LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit." In: *Computational Linguistics in the Netherlands Journal*, no. 3 (2013): 103-120.