A corpus-based study of semantic differences in translation

The case of inchoativity in Dutch

Lore Vandevoorde, Els Lefever, Koen Plevoets and Gert De Sutter
Ghent University

This paper presents a corpus-driven, statistical method for the visualization of semantic structure, thereby tackling the under-researched issue of semantics in corpus-based Translation Studies. We aim to investigate the influence of translation on the structure of semantic fields and in particular the extent to which the structure of the semantic field of inchoativity differs between original, non-translated Dutch and translated Dutch. The visualizations of the semantic field of inchoativity show that translated Dutch indeed differs from non-translated Dutch on the semantic level. Based on the exploration of the semantic fields, we furthermore formulate some hypotheses concerning the presence of the so-called universal tendencies of translation on the semantic level.

Keywords: translational hypothesis, corpus-based, statistical, semantic differences, semantic field, translation universals, inchoativity

1. Introduction

The notion of meaning has always been at the core of Translation Studies: in his paper on the Translator's Task ([1923] 2004), Walter Benjamin reflected on whether translators should express "what is meant" or rather "the way of meaning it"; in 1958, Vinay and Darbelnet formulated a number of translational procedures out of a primary concern with how meaning is (or should be) transferred through translation. More recent work on semantic preferences (Bednarek 2008) and semantic prosody (Kenny 1998; Xiao and McEnery 2006; Dam-Jensen and Korning Zethsen 2008; Korning Zethsen 2008) has raised an awareness of the relevance of corpus-based research in semantics for Translation Studies (Dam-Jensen and

Target 29:3 (2017), **388–415**. DOI 10.1075/target.15009.van
ISSN 0924–1884 / E-ISSN 1569–9986

Korning-Zethsen 2008). In addition, recent studies by Delaere, De Sutter, and Plevoets (2012) and by Diwersy, Evert, and Neumann (2014), as well as the studies presented in the edited volumes by Oakes and Ji (2012) and De Sutter, Lefer, and Delaere (2017) introduced multivariate statistics into corpus-based Translation Studies (henceforth CBTS), allowing for more empirical research on language variation in translational environments.

Despite this concern with meaning, methodological tools from empirical semantics have – to our knowledge – barely been introduced into Translation Studies proper. The question whether semantic relations between words in a specific semantic field are identical in translated and non-translated language has indeed rarely been raised within Translation Studies, possibly due to the difficulty to operationalize semantic differences in an empirical study. Strategies, for instance, to detect universal tendencies of simpler, more explicit or more conservative language use (via the comparison of grammatical structures or vocabulary between translated and original texts) do not necessarily apply to semantic phenomena. In addition, meaning is typically considered as the invariant of translation in Translation Studies, dismissing it at once as a subject of research.

The goal of this paper is therefore to put forward a methodological solution for the study of semantic differences in translated and non-translated language, making use of multivariate statistics to capture the complexity of the meaning relationships. Furthermore, a descriptive exploration of the differences between the semantic fields of inchoativity in translated and non-translated Dutch on the basis of the proposed methodology can constitute a first step towards more empirical research into semantic differences in translation. The identification of semantic differences between translated and non-translated language offers the prospect of revealing hidden cognitive processes in translation which could in turn explain some of the alleged 'universals' of translation on the cognitive level (and offer an alternative to the hitherto suggested explanations of diverging norms and conventions).

The outline of this paper is as follows. First, we propose a methodology which consists of measuring semantic distances between lexemes in different varieties of Dutch (non-translated Dutch, Dutch translated from English, and Dutch translated from French). The complexity of the meaning relationships is captured in frequency tables in which translations are defined as the variables of source-language lexemes (and vice versa). By applying Hierarchical Cluster Analysis to the obtained data sets, we generate visual representations of semantic fields of translated and non-translated Dutch. These visualizations can consequently be compared to each other so as to detect possible differences in semantic distances between the lexemes in the semantic field representing non-translated Dutch and the ones representing translated Dutch. Second, the methodology is tentatively applied to the

semantic field of inchoativity in Dutch, of which *beginnen* [to begin] is one of the most prototypical expressions in Dutch; it is used more frequently than its closest near-synonym *starten* [to start] with 291,438 hits for *beginnen* versus 23,986 for *starten* in the Dutch reference corpus SONAR (Oostdijk et al. 2013). The visualized representation of inchoativity in non-translated (source-language) Dutch (SourceDutch) is described and compared to the representations of inchoativity in translated (target-language) Dutch translated from English ($\text{TransDutch}_{\text{ENG}}$) and translated from French ($\text{TransDutch}_{\text{FR}}$). We will conclude this paper by generating a number of hypotheses about the possible presence of the so-called 'translation universals' on the semantic level. We will furthermore suggest (i) how these alleged universals can be investigated, relying on the method proposed in this paper and (ii) how theories from neurolinguistics (we will give the example of Paradis' neurolinguistics theory of bilingualism) can serve as an explanatory basis for such phenomena.

2. Methodology

Our methodological proposition consists of two parts. First, a retrieval method for candidate lexemes of a semantic field under scrutiny is developed, which has the specificity that comparable data sets can be created for translated and non-translated language. Second, the semantic relationships – revealed on the basis of the so-called 'translational equivalence' hypothesis and understood in terms of distances – are captured in so-called Semantic Vector Spaces. This enables us to visualize the semantic fields via statistical techniques for data exploration such as Hierarchical Agglomerative Clustering. On the basis of those visualizations, semantic fields of translated and non-translated language can be inspected and compared to each other.

2.1 Lexeme retrieval

Before semantic fields can be created, a set of lexemes needs to be selected as candidate members of a semantic field under study. To do so, we apply the Semantic Mirrors Method (henceforth SMM), which was originally designed by Helge Dyvik (1998, 2004, 2005) to derive large-scale semantically classified vocabularies. The SMM is based on the assumption that "semantically closely related words ought to have strongly overlapping sets of translations, and words with wide meanings ought to have a higher number of translations than words with narrow meanings" (Dyvik 2004, 311). This technique has been acknowledged for its ability "to define lexical properties as ambiguity, vagueness and synonymy, as well as lexical fields,

feature-specified hierarchies and overlap relations with these fields (e.g. prototypicality, hyponymy)" (Altenberg and Granger 2002, 29) and has indeed been applied as such, albeit mostly for intra-linguistic and contrastive purposes, and with respect to discourse markers (Mortier and Degand 2009), pragmatic markers (Aijmer and Simon-Vandenbergen 2004), and adverbs (Simon-Vandenbergen and Aijmer 2007; Simon-Vandenbergen 2013).

The retrieval of lexemes via SMM works as follows (see Figure 1): all translations in Language B of an initial polysemous 'lexeme *a* in Language A' are extracted from a sentence-aligned corpus. This set of translations is called the 'first T-image of *a* in Language B.' Then, commensurably, the translations back into Language A (the 'back-translations') of the 'first T-image' (themselves translations from *a*) are looked up. This is called the 'inverse T-image of *a* in Language A.' Finally, the initial procedure is applied a second time: the translations in Language B of the 'inverse T-image' lexemes in Language A are retrieved (this is called the 'second T-image'). Schematically, the procedure looks as follows:

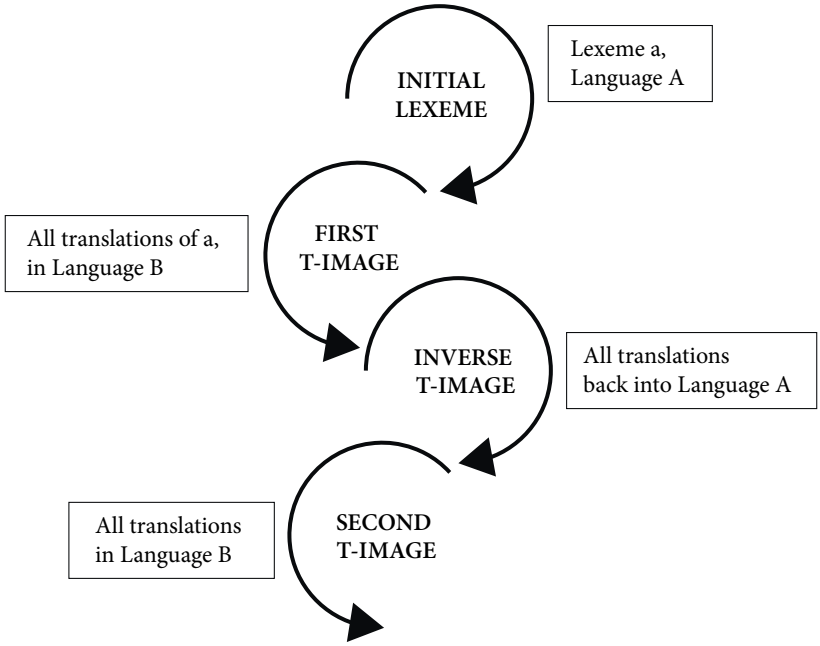

Figure 1. Semantic Mirrors Method

Dyvik's method has the potential to reveal meaning relationships on the basis of translations, as the following (fictitious) example shows (Figure 2). For the Dutch lexeme *bank* ('lexeme *a* in Language A'), we obtain a 'first T-image' in French (Language B) with *institution financière* [financial institution], *banque* [bank],

banc [bench] and *fauteuil* [armchair]. The 'inverse T-image' back into Dutch (Language A) of these French (Language B) lexemes then consists of *financiële instelling* [financial institution], *bank* [bank, seat], *sofa* [sofa], *zetel* [seat], *school* [school (of fishes)], *leunstoel* [armchair].

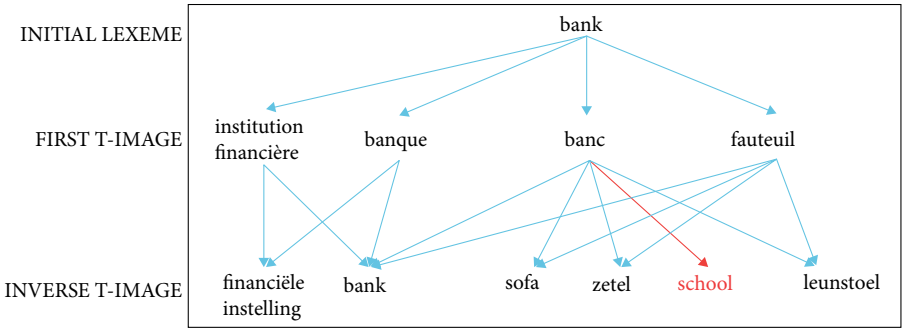

Figure 2. The SMM applied to the fictitious example of *bank*

This network of translational relationships now allows us to distinguish the different senses of *bank*. Some caution is needed, however, since at the level of the inverse T-image, we need to discard 'mutually unrelated senses' (Dyvik 2005) of the initial lexeme (*bank*). In the above example, *school* [school of fishes] is a 'mutually unrelated sense' of the initial lexeme *bank*. Indeed, "school of fishes" is not a sense of the initial lexeme *bank*. In the retrieval method, 'mutually unrelated senses' in the inverse T-image can be detected because they are only a translation of exactly one lexeme in the first T image (in our – example, *school* is only a translation of *banc* in the first T-image). As a consequence, 'mutually unrelated senses' can be avoided by selecting only those lexemes on the level of the inverse T-image that are translations of at least two first T-image lexemes. In the above example, this mechanism leads to the selection of *financiële instelling*, *bank*, *sofa*, *zetel* and *leunstoel*, whereas *school* is excluded.

In addition, we can discern 'related senses' as well as 'unrelated senses' (Dyvik 2005) of the initial lexeme. *Financiële instelling* and *bank* are 'related senses' because they are mutually connected to each other via the first T-image. The same holds for *bank*, *sofa*, *zetel* and *leunstoel*: they are related to each other because they are all mutually connected via the first T-image to the initial lexeme *bank* and at least one other lexeme in the inverse T-image. *Financiële instelling* and *sofa*, however, are 'unrelated senses' because they are, via the first T-image, only connected to the initial lexeme *bank* in the inverse T-image. Since *bank* is related to

financiële instelling, on the one hand, and to *sofa*, *zetel* and *leunstoel*, on the other hand, but *financiële instelling* is unrelated to *sofa*, *zetel* and *leunstoel*, it can be concluded that Dutch *bank* is contrastively ambiguous between the sense of *banque / institution financière* [bank, financial institution] and the sense of *banc / fauteuil* [seat].[1] Because we are interested in the different senses of an initial lexeme under scrutiny, we will include the 'related' and the 'unrelated senses' of the initial lexeme and only exclude 'mutually unrelated senses,' by applying what we call an 'overlap criterion': every lexeme selected at the level of the inverse T-image is a translation of at least two first T-image lexemes.

The SMM thus relies on the idea of 'translational equivalence' (translations can be used to reveal and disambiguate the meaning(s) of a source-language lexeme). It uses *translations* as a methodological instrument to generate semantic information but it does not as such investigate the phenomenon of translation, nor does it inquire into the possible influence of translation on the obtained semantic information. Within the scope of this research – and within CBTS more generally – the status of translations is quite different; for they are also the *subject* of research (rather than the *instrument*). If we want to use Dyvik's technique to investigate translational phenomena, we will need to somewhat *extend* the technique. These two extensions, presented in the following Section 2.2 and 2.3 will lead to the development of the Extended Semantic Mirrors Method (SMM++).

2.2 From lexeme retrieval to the compilation of comparable data sets for translated and non-translated language

How can the SMM now be adapted in such a way that it can serve the pursued methodological goal, namely, to create visual representations of semantic fields of translated and non-translated language?

The SMM uses translations as a means to reveal semantic relations. By using only translations, and by going back-and-forth between languages, a semantic hierarchy can be created, as is shown by the *bank* example. However, the *nature* of the data (translated or non-translated) is not taken into account within the SMM (unsurprisingly, as it was initially conceived as a retrieval technique). If we take a look at Table 1 below, we see that Language A is a source language in the first and the second T-image and a target language in the inverse T-image:

1. For clarity's sake, we did not include the second T-image in the example. In the SMM, the second T-image is created to arrive at a network of related lexemes for Language B, parallel to the network created for Language A in the inverse T-image.

Table 1. Nature of the data within the SMM

Step of SMT	Source language	Target language	
Initial lexeme a	Dutch		
First T-image	Dutch	English	
Inverse T-image	English	Dutch	translated/target Dutch
Second T-image	Dutch	English	original/source Dutch

This implies that the data sets which are yielded by the different steps of the SMM are different in translational nature: the data set retrieved at the level of the inverse T-image can be used to analyze translated (target) Dutch, whereas the data set retrieved at the level of the second T-image can be utilized to analyze non-translated (original/source) Dutch.

Our first 'extension' consists in a differentiation between sets of retrieved data within the different steps of the SMM based on their *translational status* (source or target language). Instead of using the second T-image to make a contrastive comparison (like Dyvik), we assign a new role to this step of the SMM, based on the translational status of the data. This is a necessary first step to make the data obtained via the SMM usable for Translation Studies research: data retrieved at the level of the inverse T-image represent translated language, data retrieved at the level of the second T-image represent non-translated language.

2.3 Visualization of data sets

The next step is to arrive at comparable visualizations of those sets of lexemes. The information which has so far been obtained only gives the researcher 'paired sets of lexemes' (each time a source-language lexeme with its corresponding target-language lexeme). Within the original SMM, hierarchical patterns are indeed "only based on overlap relations among t-images" (Dyvik 1998, 73) and are obtained by ranking the lexemes "independently of frequency of occurrence" (ibid.). The degree of semantic similarity between the lexemes in the created hierarchy is only based on the number of overlapping translations while frequency information is excluded. Table 2 shows a fictitious example of the translational relation in the inverse T-image of Dutch *bank* with French as a pivot language. Based on this information, the centrality of *bank* in a field with *bank*, *financiële instelling*, *sofa* and *leunstoel* can be deduced from the fact that *bank* is a translation of all three French lexemes *banque*, *banc* and *fauteuil*.

Table 2. Overlapping translations of *bank* (fictitious) in the inverse T-image

is translated as →	**bank[nl]**	**financiële instelling**[nl]	**sofa**[nl]	**leunstoel**[nl]
banque[fr]	X	X	Ø	Ø
banc[fr]	X	Ø	X	X
fauteuil[fr]	X	Ø	X	X

The degree of similarity can be introduced into the rationale in a statistically meaningful way by collecting occurrence counts of words with other words/features in a frequency table (which is what is typically done in distributional semantics). Frequencies indicate the strength of certain relations: they will tell us which patterns are important. They can be considered as multivariate data and are "typically represented in a matrix form with rows holding the units and columns holding the variables" (Jenset and McGillivray 2012, 302). As a consequence, such frequency tables can be thought to represent translated language when the translated lexemes are represented as rows with their source-language lexemes as column variables. They can represent non-translated language when the non-translated (source-language) lexemes are represented as rows with their translations as column variables. The integration of frequency information in the SMM is our second major adaptation of the SMM. If we apply the idea to the previously given fictitious example of *bank*, the result then looks as follows for non-translated (original/source) language *bank* (Table 3) and translated (target) language *bank* (Table 4):

Table 3. Frequency table for original *bank* – second T-image (fictitious)

is translated n times as →	**banque**[fr]	**banc**[fr]	**fauteuil**[fr]
bank[nl]	231	61	45
financiële instelling[nl]	178	0	0
sofa[nl]	0	124	32
leunstoel[nl]	0	27	76

Table 4. Frequency table for translated *bank* – inverse T-image (fictitious)

is n times a translation of →	**banque**[fr]	**banc**[fr]	**fauteuil**[fr]
bank[nl]	230	61	45
financiële instelling[nl]	121	0	0
sofa[nl]	0	98	32
leunstoel[nl]	0	67	43

The use of frequency tables allows us to carry out advanced statistical techniques upon our data sets, and opens the way to statistical visualization techniques such as Correspondence Analysis (Greenacre 2007; Lebart, Salem, and Berry 1998) and Hierarchical Cluster Analysis (Baayen 2008, 138; Gries 2013, 336), a technique that will allow us to visually represent the similarities and differences between the sets of lexemes. Previous research in Contrastive Linguistics has shown that Hierarchical Cluster Analysis is an excellent tool for the evaluation of corpus-based, lexico-semantic analyses (Divjak and Gries 2009; Gries 2012; Divjak and Fieller 2014).

3. The semantic field of inchoativity in translated and non-translated language

A practical prerequisite to carry out the proposed method is that the researcher needs to have access to a parallel corpus, which is preferably at least sentence-aligned. If the corpus is word-aligned, the researcher can work in the most optimal circumstances (but word-alignment can be carried out manually or (semi-) automatically on the parallel sentences under investigation). The data for this study were extracted from the Dutch Parallel Corpus (DPC), a ten-million-word sentence-aligned, both parallel and comparable corpus. It is balanced with respect to five text types (external communication, journalistic texts, instructive texts, administrative text, fictional and non-fictional literature) and four translation directions (Dutch to French, French to Dutch, Dutch to English, and English to Dutch) (Macken, De Clercq, and Paulussen 2011, 376–378). Although the DPC has numerous advantages, it is not word-aligned. In order to speed up the manual annotation process, we used the statistical word alignment model GIZA++. All the automatically aligned data were subsequently manually controlled and validated. At every level of SMM, invalid alignments were eliminated from the data.

3.1 Retrieval

In order to generate the semantic field of inchoativity, we selected an initial lexeme in Dutch, which we consider as the most prototypical expression of inchoativity in Dutch, namely *beginnen* [to begin] (corresponding to 'lexeme *a* in Language A'). The choice for inchoativity offers a number of advantages: (i) we expect high corpus frequencies of lexical items expressing inchoativity, which will facilitate statistical processing; (ii) for two central Dutch expressions of inchoativity, namely, *beginnen* and *starten*, close cognate translations are available in English (*to begin* and *to start*) but this is not the case in French (a particularity which can possibly offer interesting contrastive perspectives, for instance, about the impact of close cognates on the structure of semantic fields of translated language); (iii) the meaning differences between the expressions of inchoativity are expected to be (very) fine-grained (Schmid 1996). Inchoativity is therefore a compelling test case when one is interested in revealing meaning differences. Admittedly, numerous other cases could be studied on the basis of countless other grounds, but the advantages enumerated above make the case of inchoativity an interesting point of departure for our study.

The French (language B) translations of the initial lexeme *beginnen* were manually checked (n = 398) resulting in a set of 75 unique French translations (the first T-image, corresponding to all translations of *a* in Language B). We subsequently selected those French lexemes that were attested at least 3 times[2] in the corpus as translations of the initial lexeme *beginnen*. This yielded a first T-image of 19 different French lexemes. Next, the 19 selected first T-image lexemes were queried from the corpus as source-language lexemes (n = 1706). Their translations back into Dutch were manually checked, and the same frequency threshold of 3 attestations and a 'minimal overlap criterion' (every selected lexeme is the translation of at least 2 source-language lexemes) was applied. In this way, 'mutually unrelated

2. Carrying out the SMM++ with a frequency threshold of 2 would have resulted in the following 9 lexemes to be added to this list: *aangaan* [to enter into], *aanvatten* [to commence], *begin-* [begin-], *doen* [to do], *lanceren* [to launch], *maken* [to make], *nemen* [to take], *sinds* [since], *start-* [start-]. Admittedly, a frequency threshold of two observations would have been the better conceptual choice. Indeed, the lower the frequency threshold, the larger the number of lexemes involved and the more information could be gained from the analysis. The drawback of a lower frequency threshold lies in the large amount of additional manual annotation work that this would involve, a task which could not be completed within the ambit of this study. This difficulty could be overcome by relying on automatic word alignment or word aligned parallel corpora. While the latter are not available for the languages we are investigating, the former task was tentatively carried out with GIZA++. However, the corpus size of the DPC appeared too small, so that many automatic word alignments were incorrect and needed consistent and thorough manual validation.

senses' of the initial lexeme are excluded here. This step resulted in an inverse T-image of 36 unique Dutch lexemes (the inverse T-image, corresponding to all translations back into Language A). Finally, the French translations of the inverse T-image were again queried from the corpus resulting in the second T-image (all translations in Language B) (n = 1490).

We repeated this procedure with English as a Language B. The English translations of *beginnen* were manually checked (n = 336) and after applying the same frequency threshold of 3, a first T-image of 9 different English lexemes was yielded. The first T-image lexemes were inversely queried as source-language lexemes (n = 1029) and their translations back into Dutch resulted in an inverse T-image of 24 unique Dutch lexemes. Finally, the English translations of the inverse T-image were again queried from the corpus to create the second T-image (n = 1117). In order to be able to compare the data set for SourceDutch with those for $\text{TransDutch}_{\text{ENG}}$ and $\text{TransDutch}_{\text{FR}}$, we selected the common lexemes amongst those data sets. In total, 16[3] lexemes were independently selected by the mirroring of both $\text{beginnen}_{\text{ENG}}$ and $\text{beginnen}_{\text{FR}}$. These 16 Dutch lexemes are: *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], *worden* [to become].

Table 5. Absolute frequencies of the Semantic Mirroring of *beginnen*

	Source language	**Target language**	**Total *n***	**Selected translations**
$\text{beginnen}_{\text{FR}}$ first T-image	Dutch	French	398	19
$\text{beginnen}_{\text{FR}}$ inverse T-image	French	Dutch	1706	36
$\text{beginnen}_{\text{FR}}$ second T-image	Dutch	French	1822	36
$\text{beginnen}_{\text{ENG}}$ first T-image	Dutch	English	336	9
$\text{beginnen}_{\text{ENG}}$ inverse T-image	English	Dutch	1029	24
$\text{beginnen}_{\text{ENG}}$ second T-image	Dutch	English	1117	24

3.2 Visualization

For the first visualization – that of the semantic field of SourceDutch (Figure 3) – we use the (source-language) frequencies of the second T-image of both datasets

3. In fact, 17 lexemes were selected, but *vanaf* [as from] was discarded on the basis of a preliminary inspection of the data on the basis of the Correspondence Analysis because of its outlying position in the spatial map.

(beginnen$_{FR}$ and beginnen$_{ENG}$). The SourceDutch data set contains the data of two semantic mirrors (obtained in the second T-image of beginnen$_{FR}$ and beginnen$_{ENG}$). In this way, semantic features from two distinct languages are combined and the neutralization of any possible influence of the semantic features (translations) on the visualization of SourceDutch is maximized. For the visualization of the semantic fields representing translated language (TransDutch$_{FR}$ and TransDutch$_{ENG}$), we use the (target-language) frequencies of the inverse T-image and we apply the same statistical techniques.

The statistical technique of Hierarchical Cluster Analysis is now applied to each data set. All our analyses are carried out with the open-source statistical software R (R Core Team 2014). While most analyses can be carried out using existing packages in R, we use the svs-package (Plevoets 2015), which contains "various tools for semantic vector spaces" for a number of analyses. A Hierarchical Cluster Analysis is a "graphical representation of a matrix of distances […] where the objects are joined together in a hierarchical fashion from the closest, that is most similar, to the furthest apart, that is the most different" (Greenacre 2013, 89). This means that elements within a distinct cluster of the visualization are as similar as possible, whereas the similarity distance between different clusters is as large as possible.

The Hierarchical Cluster Analysis was in fact not carried out on the raw frequencies of the source- or target-language lexemes, but on the output of a Correspondence Analysis (Baayen 2008, 128; Greenacre 2007; Lebart, Salem, and Berry 1998), which offers us two advantages. First, Correspondence Analysis is an exploratory graphical technique, which represents the data in a low-dimensional space, thus allowing for the inspection of patterns in the data. Second, the visual representation of Correspondence Analysis involves dimension reduction, which means that noisy patterns can be filtered out from informative patterns, the latter being typically represented on the first few axes with a large variance (or 'inertia'), whereas the noisy patterns are represented on the later axes with a small variance.

Both advantages of Correspondence Analysis also provide subsequent reasons to apply Hierarchical Cluster Analysis. First, not all patterns in the spatial representation are clear-cut. We can therefore use the coordinates of the lexemes in the Correspondence Analysis as input for the Hierarchical Cluster Analysis in order to make our assertions about the semantic spaces of translated and non-translated language more objective. Second, the distribution of inertia over the latent dimensions of the CA can be visualized in a so-called scree plot which indicates how many dimensions are needed to reach a threshold, for instance, 80%. A two-dimensional plot only visualizes the first two dimensions of the CA, and can, as a consequence, hide important information.

The Euclidean distance between all the coordinates from the Correspondence Analysis was computed and the points were clustered by means of Ward's Minimum Variance Method.[4] In order to solidify the results of the Hierarchical Cluster Analysis, we furthermore obtained statistical significances by means of the R package pvclust (Suzuki and Shimodaira, 2006). For each edge in a cluster tree (that is, each point where two branches join), pvclust enables us to compute both an "approximately unbiased p-value" (depicted in red) and a "bootstrap probability" (depicted in blue). Those computations are based on bootstrapping, which means that the data are resampled (with replacement) a number of times, in order to see how much they (in this case, the clusters) vary. The number of bootstrap replications is usually fairly high, and we have chosen for 3000 replications. The literature on pvclust itself recommends the approximately unbiased p-value (abbreviated as *au*) as the more accurate measure of the statistical significance of a cluster. A significant cluster is then the *highest* edge in a tree which has a p-value above a certain confidence level. In line with common practice, we use a confidence level of 95%.

4. Results

In this section, we describe the created cluster trees or dendrograms that were yielded via the statistical technique of Hierarchical Cluster Analysis. We first describe the semantic field of SourceDutch (Figure 3), which we will subsequently compare to the semantic fields of translated Dutch (Figure 4: $\text{TransDutch}_{\text{ENG}}$, and Figure 5: $\text{TransDutch}_{\text{FR}}$).

4.1 The semantic field of inchoativity for SourceDutch

We will now provide an interpretation of the visualization representing a semantic field of *beginnen* / inchoativity for SourceDutch (Figure 3). This interpretation will be used to determine a meta-label for each cluster so as to name the specific meaning distinction revealed by that cluster. The assigned meta-labels should be

4. Vandevoorde (2016) compares various combinations of distance measures (Euclidean and Canberra) and amalgamation rules (Average, Complete, Ward's) on different spatial maps (HCA on raw data, HCA on the output of a Correspondence Analysis and HCA on the output of a Latent Semantic Analysis) in order to see which combination works best. The overall strength of the clustering structure of these various combinations is assessed on the basis of the agglomerative coefficient, the amount of chaining effect and the p-values. Based on this assessment, the combination of Euclidean Distance and Ward's Minimum Variance Method on the output of a Correspondence Analysis was selected.

understood as a post-hoc interpretative tool, applied to enhance our understanding of the rendered dendrograms. Information from three types of sources will be used: (i) corpus examples from the DPC containing the lexemes which make up a cluster, (ii) attestations in reference works, and (iii) information from the lexical database Cornetto (Vossen et al. 2008).

We consider cluster n°3 as the cluster representing the idea of GENERAL ONSET. This cluster holds *beginnen*, which was selected as the initial lexeme of the retrieval task and is considered as the most prototypical expression of inchoativity. We see that the cluster contains three different sub-nodes, one with *opstarten* [to start up], and two other, interrelated sub-nodes; one with *starten* [to start] and *van start gaan* [to take off] and another one with *beginnen* [to begin] and *gaan* [to go]. In our opinion, these latter two interrelated sub-nodes indicate an additional meaning-distinction within the meaning-distinction indicated by cluster n°3. Next to *beginnen*, *starten* is also a typical expression of inchoativity and the two are often considered as near-synonyms (Schmid 1996, 223). Divjak and Gries (2009) – based on research by Biber et al. (1999) and Schmid (1993), following Quirk et al. (1985) – conclude the following for the English phasal verbs *to start* and *to begin*:

> Begin then gives a view into the *state after onset* of the action: it expresses modality/intentionality and refers to later states of affairs. It typically applies to cognitive-emotive events and non-perceivable things. Start, on the other hand, focuses on the *actual action*, the actual beginning, the very moment of transition from non-action to action. It is dynamic and applies to visible change and actions.
>
> (Divjak and Gries 2009, 279; our emphasis)

The subdivision observed in our (Dutch) results into verbs formally related to *starten* [to start] such as *opstarten* [to start up] on the one hand (henceforth ACTION verbs), and verbs formally related to *beginnen* [to begin] (henceforth STATE AFTER ONSET verbs) on the other hand, thus corroborates the distinction made by Divjak and Gries for English *to begin* and *to start*. The attested distinction between *to begin* and *to start* seems to hold for Dutch *beginnen* and *starten* too. Cornetto (Vossen et al. 2008) defines *gaan* [to go] – which is at first sight somewhat oddly clustered with *beginnen* – as "*beginnen iets te doen*" [to begin to do something], and *beginnen* as "*iets gaan doen*" [to go and do something]. The definitional relation indicated by Cornetto thus seems to underpin the semantic relationship indicated by the clustering of *beginnen* and *gaan*. In addition, according to the *Algemene Nederlandse Spraakkunst* [General Dutch Grammar] (Haeseryn et al. 2012), the first of two subtypes of *gaan* "without the meaning of motion" is the subtype where *gaan* has the meaning of "'*(geleidelijk) overgaan tot', 'beginnen te' (inchoatief aspect)*" [(gradually) move on to, to begin to (inchoative aspect)]. The relatedness between *starten* and *beginnen* is also further substantiated by the

definitions of *starten* in Cornetto: (i) "*beginnen van iets (niet-causatief)*" [beginning of something (non-causative)], (ii) "*doen beginnen (causatief)*" [to make begin (causative)], and (iii) "*(van apparaten) beginnen te functioneren*" [(of devices) to begin to function], all of which bear *beginnen* in their Dutch definition. In sum, we decide to assign the label of GENERAL ONSET to cluster n°3. We furthermore discern an additional meaning distinction within this cluster between ACTION verbs (to which we will assign the label ACTION) and STATE AFTER ONSET verbs (which will be labeled as STATE AFTER ONSET).

Cluster n°2 contains *begin* and *start* – which are the nominal derivatives of *beginnen* and *starten* – as well as *aanvang*, which is again a noun, but differs from *begin* and *start* in that it belongs to a more formal register. Although the majority of the lexemes in the dendrogram are verbs, there are indeed three nouns represented, which are now grouped together into one cluster. A possible explanation for the separate clustering of the nouns and verbs in our analysis goes as follows: a nominal derivative such as *begin* and its 'root' verb *beginnen* appear in different syntactic contexts but are likely to appear in similar lexical environments. Since our analysis can be considered as a translational analysis, which uses translation to lay bare meaning, it seems plausible that the syntactic environment of a sentence is more likely to primarily impose choice of word class (e.g., a noun is more likely to be translated by a noun, and a verb by a verb), which could explain why our translational method favors a word-class dependent clustering of lexemes. Based on the previous reflection, we decide to use GENERAL ONSET (NOUN) as the metalabel for cluster n°2.

Cluster n°1 holds the verbs *oprichten* [to set up, to establish] and *opzetten* [to set up]. Within Cornetto *oprichten* is defined as *opzetten*. We consequently consider them as near-synonyms. In Cornetto, *oprichten* is associated with the setting up of an association, a party, a school; whereas *opzetten* is associated with the setting up of a project, an activity, a bank, a company, a business. Corpus examples (1 and 2) from the DPC show that *oprichten* can, just as *opzetten*, be used in business-like contexts:

(1) TARGET: *In 2000* ***zetten*** *de twee bedrijven een joint venture* ***op*** *in Turkije*
SOURCE: In 2000 the two companies **set up** a joint venture together in Turkey
(dpc-arc-002048-en; our emphasis)

(2) TARGET: *Company1 versterkt zijn positie in het Oosten en* ***richt*** *filialen* ***op*** *in Australië en Taiwan.*
SOURCE: Company1 strengthens its position in the east and **starts up** subsidiaries in Australia and Taiwan.
(dpc-bco-002345-en; our emphasis)

What seems to distinguish this cluster from the cluster of GENERAL ONSET is that *opzetten* and *oprichten* appear to indicate a specific type of action, related to the setting up of a project, a business, a company, etc. We will therefore add the label SPECIFIC ACTION to cluster n°1.

The lexemes *komen* [to come], *krijgen* [to get], *worden* [to become] in cluster n°5 share the semantic characteristic that their inchoative aspect is non-lexicalized. By this we mean that these verbs' potential to express inchoativity is not directly apparent from the verbs themselves but that these verbs receive their inchoative value from the context they are used in (compared to, for instance, *beginnen*, in which the inchoative aspect is lexicalized, and hence, directly apparent irrespective of the context it is used in). Witness the following example (note that, in this Example (3), the inchoative aspect is made explicit by its translation):

(3) SOURCE: *SteelUser is er* ***gekomen*** *om onze klanten het leven een stuk aangenamer en eenvoudiger te maken […]*
TARGET: SteelUser was **set up** to make life simpler and more comfortable for our clients […] (dpc-arc-002053-nl; our emphasis)

In Cornetto, the inchoative aspect of the three verbs is implicitly present in one of the definitions of *komen*, namely, "*beginnen te spreken*" [start to speak], of *krijgen*, namely, "*in een situatie terechtkomen*" [to find oneself in a situation], and in the examples provided by Cornetto for the copulative verb *worden* [to become], "*boos/ziek/misselijk worden*" [to become angry/ill/nauseated]. The meta-label chosen for this cluster is NON-LEXICALIZED INCHOATIVITY.

Ontstaan [to come into being] and *openen* [to open] make up cluster n°4. *Ontstaan* is defined as "*tot stand komen*" [to come about] in Cornetto. *Openen*, in its inchoative meaning, is defined as (i) "*laten beginnen*" [to let begin] when its semantic type is 'action' ("describing an action usually controlled by the subject of the verb"), and as (ii) "*opengaan*" [to open] when its semantic type is 'process' ("not initiated by an actor capable of acting with volition"). The examples in Cornetto indicate that *ontstaan* is often used to indicate the coming into being of abstract processes such as fights or quarrels ("*ruzie/onenigheid ontstaat*" [a fight/a disagreement arises]), or either for the coming into being of natural phenomena such as mountains or rivers ("*een gebergte ontstaat*" [a mountain chain comes into being]; "*een rivier ontstaat uit een bron*" [a river originates from a source]). *Openen* is used to introduce the beginning of an event, either as an 'action' (controlled by the subject of the verb), as in "*een symposium openen*" [to open a symposium] or as a 'process' (not initiated by an actor capable of acting with volition), as in "*het symposium opent*" [the symposium begins]. Although this is not explicitly mentioned in Cornetto, the corpus furthermore (Example (4)) shows that *openen* can, just as *ontstaan*, refer to abstract processes, such as the coming into being of a right:

(4) *Ik kan het recht* ***openen*** *op een tegemoetkoming omdat ik tot 21 jaar de verhoogde kinderbijslag genoot* [I can **open** the right on subsidy because I received increased family allowance until the age of 21]
(dpc-fsz-001052-nl; our emphasis).

The meaning distinction of the clustering of *openen* and *ontstaan* will tentatively be captured with the meta-label ONSET OF ABSTRACT PROCESSES, which seems to be the common denominator of the two verbs.

Finally, cluster n°6 is a singleton cluster, containing the adverb *eerst* [firstly], which presents a clear inchoative meaning. Again, just as nouns were not clustering with verbs, the only adverb is our set of candidate lexemes does not cluster with any other lexemes, further substantiating the previously made observation that our method favors word-class dependent clustering.

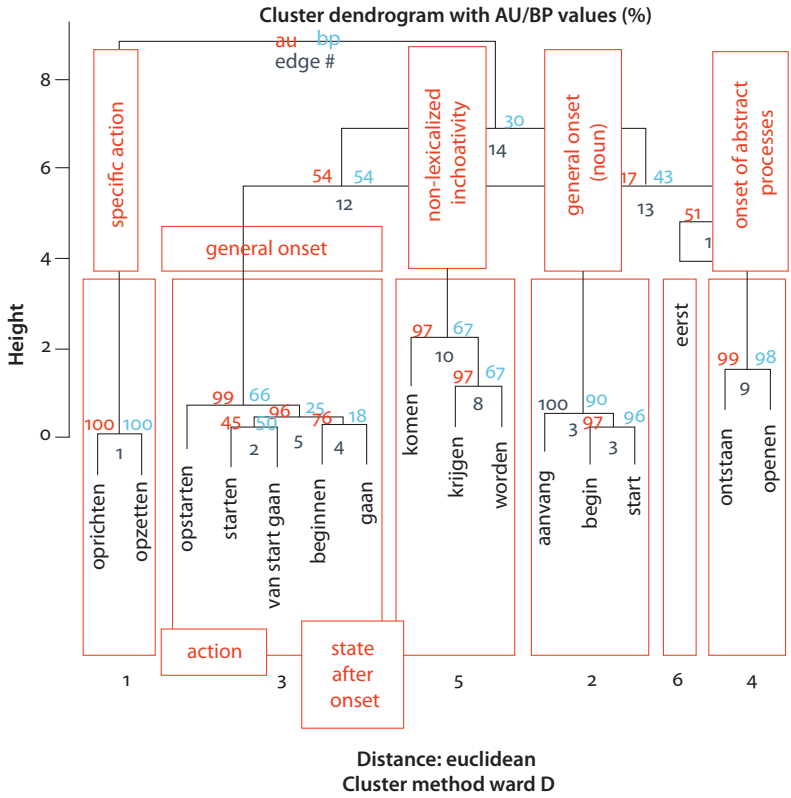

Figure 3. SourceDutch semantic field of inchoativity

4.2 The semantic field of inchoativity for TransDutch$_{ENG}$

We now provide an interpretation of a semantic field of *beginnen* / inchoativity for TransDutch$_{ENG}$ (Figure 4). The specific meaning distinctions determined for SourceDutch will be used as a point of reference to interpret the field of TransDutch$_{ENG}$. We attempt to assign the meta-labels that were chosen on the basis of the SourceDutch field to the field of TransDutch$_{ENG}$.

We consider cluster n°3 as the cluster of GENERAL ONSET. Parallel to SourceDutch, this is substantiated by the presence of the initial lexeme *beginnen* (considered as the most prototypical expression of inchoativity). We notice that the cluster of GENERAL ONSET has become larger compared to SourceDutch: *eerst* – which held a peripheral position in SourceDutch – is now part of the cluster of GENERAL ONSET, as well as *krijgen* and *worden*, labeled as NON-LEXICALIZED INCHOATIVITY in SourceDutch. This implies that more peripheral expressions of inchoativity, as well as expressions where inchoativity is non-lexicalized, are used more prominently to express inchoativity in TransDutch$_{ENG}$, compared to SourceDutch.

Just as we did for SourceDutch, we will now further inspect the different subnodes of the cluster of GENERAL ONSET to see whether the same meaning distinction between ACTION and STATE AFTER ONSET is also present in TransDutch$_{ENG}$. We observe three sub-nodes, one higher subnode with *eerst* and *van start gaan* and two lower sub-nodes of which one with *beginnen* and *krijgen* and a second one with *starten*, *gaan* and *worden*. Whereas for SourceDutch, the subnodes of the cluster of GENERAL ONSET clearly laid bare a division between ACTION and STATE AFTER ONSET, we see that this is no longer the case in TransDutch$_{ENG}$ (e.g., *gaan* is clustered with *starten*). As a consequence, it seems that within the cluster of GENERAL ONSET of TransDutch$_{ENG}$, the emphasis is on the wider relatedness between the verbs rather than on the division between ACTION and STATE AFTER ONSET.

Cluster n°4 is a somewhat odd, new cluster. Since the cluster of GENERAL ONSET contains the ACTION verbs as well as verbs of NON-LEXICALIZED INCHOATIVITY, one would have expected *opstarten* and *komen* in the cluster of GENERAL ONSET too. Recall that in SourceDutch *opstarten* already formed a significant sub-node within the cluster of GENERAL ONSET. This distinction now seems to be emphasized in TransDutch$_{ENG}$ by the separate clustering of *opstarten*.

Cluster n°2 contains *aanvang* and *start*. Based on statistical significance, cluster n°6 – a singleton cluster with *begin* – is connected in a higher (less significant) node to *aanvang* and *start*. The word-class dependent clustering observed for SourceDutch is maintained.

The clustering within clusters n°1 (*oprichten* with *opzetten*) and n°5 (*ontstaan* with *openen*) have remained unaltered with respect to their corresponding clusters in SourceDutch.

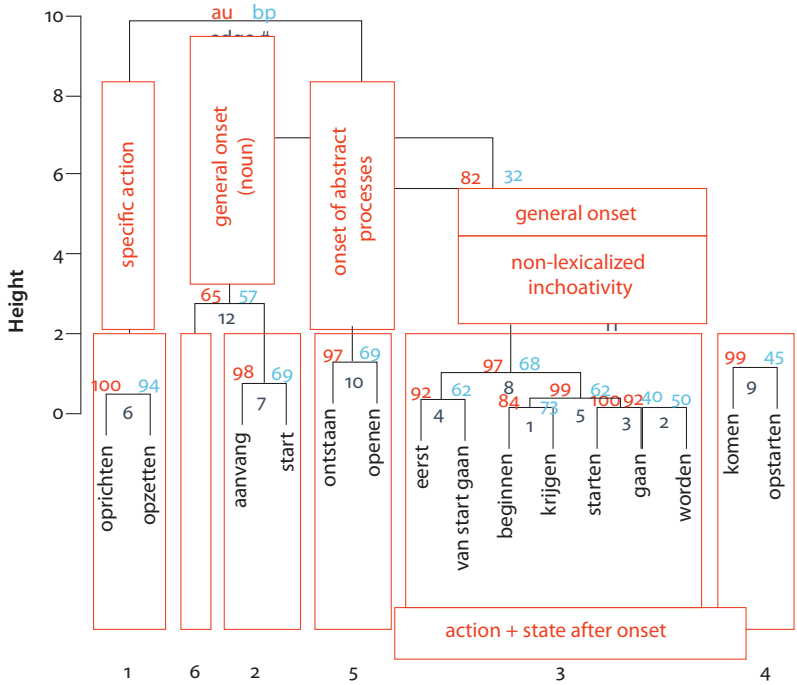

Figure 4. TransDutch$_{ENG}$ semantic field of inchoativity

4.3 The semantic field of inchoativity for TransDutch$_{FR}$

Thirdly, we provide an interpretation of a semantic field of *beginnen* / inchoativity for TransDutch$_{FR}$ (Figure 5). The specific meaning distinctions determined for SourceDutch will again be used as a point of reference to interpret the field of TransDutch$_{FR}$. Just as we did for TransDutch$_{ENG}$, we will attempt to assign these meta-labels to the field of TransDutch$_{FR}$.

We consider cluster n°4 as the cluster representing the idea of GENERAL ONSET. The cluster contains *beginnen*, leading to the assumption that this cluster contains the most prototypical expressions of inchoativity. Parallel to TransDutch$_{ENG}$, the number of lexemes in the cluster of GENERAL ONSET has increased compared to SourceDutch (5 lexemes in the cluster of GENERAL ONSET of SourceDutch, 7 for TransDutch$_{ENG}$, and 6 for TransDutch$_{FR}$). Just as for TransDutch$_{ENG}$, *eerst* – which held a more peripheral position in SourceDutch – and the verbs *komen*, *krijgen* and *worden* (NON-LEXICALIZED INCHOATIVITY) are now also part of the

cluster of GENERAL ONSET. We can conclude that for both the TransDutch fields, more peripheral expressions of inchoativity as well as verbs which do not lexicalize inchoativity are used more prominently to express inchoativity compared to SourceDutch. Within the cluster of GENERAL ONSET, we discern the meaning distinctions STATE AFTER ONSET and NON-LEXICALIZED INCHOATIVITY. An important difference with SourceDutch and $\text{TransDutch}_{\text{ENG}}$ is that the cluster of GENERAL ONSET of $\text{TransDutch}_{\text{FR}}$ no longer contains any of the ACTION verbs but only STATE AFTER ONSET verbs (*beginnen* and *gaan*). Recall that in SourceDutch, ACTION and STATE AFTER ONSET verbs formed different meaning distinctions in the cluster of GENERAL ONSET, and that for $\text{TransDutch}_{\text{ENG}}$ this distinction was still present in the cluster of GENERAL ONSET although less clear.

Cluster n°3 contains two significant sub-nodes, one with *starten* and *van start gaan*, the other one with *oprichten, opzetten, opstarten*. Within cluster n°3 we discern two meaning distinctions: SPECIFIC ACTION (*oprichten* and *opzetten*) as well all the ACTION (*starten* and *van start gaan*). In $\text{TransDutch}_{\text{FR}}$, the distinction between ACTION and STATE AFTER ONSET verbs is marked more clearly, compared to both SourceDutch and $\text{TransDutch}_{\text{ENG}}$: the clustering of the ACTION verbs with the verbs of SPECIFIC ACTION seems to emphasize the dynamic nature of these verbs. We furthermore see that *opstarten* (which formed a separate sub-node in the cluster of GENERAL ONSET of SourceDutch and a separate cluster in $\text{TransDutch}_{\text{ENG}}$) is now part of the sub-node with *oprichten* and *opzetten*, emphasizing the relatedness of *opstarten* to the specific contexts in which *opzetten* and *oprichten* are used, namely, business-like activities. These contexts are confirmed for *opstarten* by both examples in Cornetto "*een nieuw bedrijf in de V.S. opstarten*" [to start up a new company in the U.S.] and by corpus examples (5 and 6) from the DPC:

(5) SOURCE: *Toen de buizenfabriek van Kimanis in augustus* ***opgestart*** *werd […]*
TARGET: When the pipe manufacturing facility in Kimanis was **started up** in August […] (dpc-arc-002049-nl, our emphasis)

(6) SOURCE: *In sterk ontwikkelde economieën worden bedrijven vooral* ***opgestart*** *wegens een (markt)opportuniteit.*
TARGET: Companies in highly developed economies are usually **started up** on the basis of a (market) opportunity.
(dpc-vla-001161-nl, our emphasis).

To conclude, the lexemes that were covered under the meta-label GENERAL ONSET are now spread over two clusters according to the additional meaning distinction ACTION / STATE AFTER ONSET. Both cluster n°3 and cluster n°4 also contain an additional meta-label, namely, SPECIFIC ACTION for cluster n°3 and NON-LEXICALIZED INCHOATIVITY for cluster n°4.

Cluster n°1 contains the nouns *start*, *aanvang* and *begin*. Just as in SourceDutch, all three nouns are now again part of one, significant cluster. Note that the only three nouns in the set of lexemes are again clustered together, once more confirming the word-class dependent clustering. The situation is different from that for TransDutch$_{ENG}$, where *begin* formed a new, singleton cluster, and *aanvang* and *start* were clustered together.

Finally, cluster n°2 contains *ontstaan* and *openen*. This is the only cluster that has remained unaltered throughout SourceDutch, TransDutch$_{FR}$ and TransDutch$_{ENG}$.

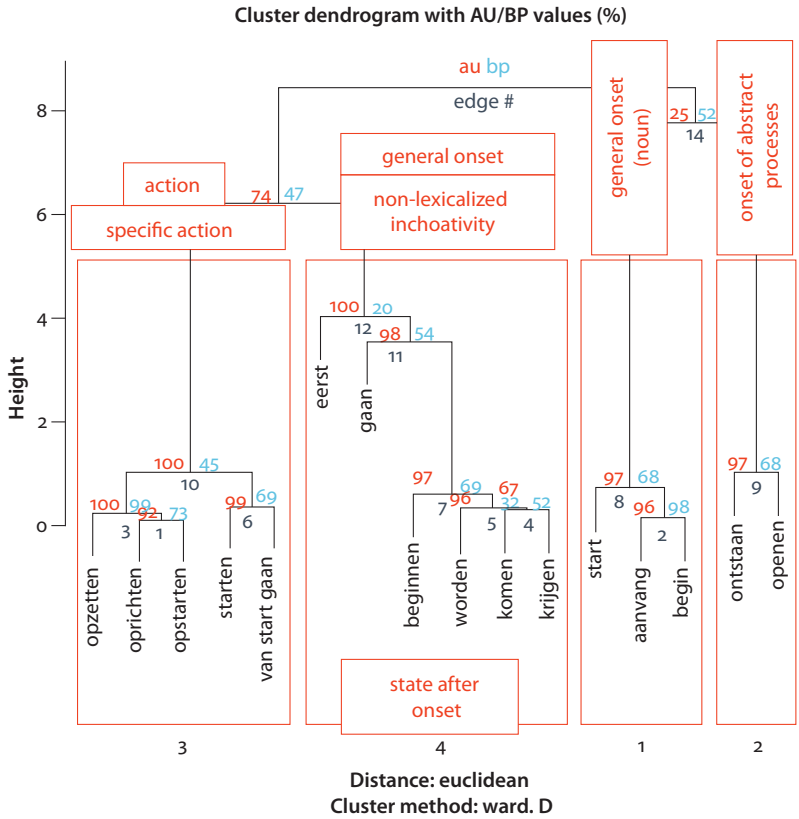

Figure 5. TransDutch$_{FR}$ semantic field of inchoativity

4.4 Provisional conclusions

The semantic fields of inchoativity in translated and non-translated Dutch show a number of differences. First, in both TransDutch$_{ENG}$ and TransDutch$_{FR}$ the cluster of GENERAL ONSET holds more lexemes than in SourceDutch: lexemes which

were labeled as NON-LEXICALIZED INCHOATIVITY in SourceDutch, and the more peripheral lexeme *eerst* in SourceDutch becomes part of the cluster of GENERAL ONSET in both TransDutch fields. The more fine-grained meaning distinctions in SourceDutch (the distinct clustering of NON-LEXICALIZED INCHOATIVITY and *eerst*) are in fact 'absorbed' into the cluster of GENERAL ONSET in TransDutch. It could hence be concluded that in both fields representing translated language, fewer fine-grained meaning distinctions can be discerned compared to the field of non-translated language.

Second, the separate clustering of *opstarten* and *begin* in $\text{TransDutch}_{\text{ENG}}$ as well as the unclear separation between ACTION and STATE AFTER ONSET in that same field lead to the unexpected conclusion that the $\text{TransDutch}_{\text{ENG}}$ field appears to deviate more from SourceDutch than $\text{TransDutch}_{\text{FR}}$. Given the fact that both Dutch and English are Germanic languages whereas French is a Romance language, one would have expected $\text{TransDutch}_{\text{ENG}}$ to be more similar to the SourceDutch semantic field (more similar languages would then be thought to give fewer 'possibilities' for deviation away from the structure of the SourceDutch field).

Thirdly, and in connection to the previous point, we observed that in $\text{TransDutch}_{\text{FR}}$, the division between ACTION and STATE AFTER ONSET is more clearly marked (they make up separate clusters) compared to $\text{TransDutch}_{\text{ENG}}$. Again, this result is quite unexpected: since the division between ACTION and STATE AFTER ONSET also exists in English, one would have expected – assuming there to be an effect of source-language influence – to see a clearer division between the two in $\text{TransDutch}_{\text{ENG}}$ compared to $\text{TransDutch}_{\text{FR}}$.

In addition, lexemes expressing ACTION and SPECIFIC ACTION are clustered together in $\text{TransDutch}_{\text{FR}}$, which seems to emphasize the dynamic nature of the cluster in contrast to the cluster of GENERAL ONSET (which also holds STATE AFTER ONSET).

5. Discussion

In this paper, we have put forward a method for the visualization of semantic fields in translated and non-translated Dutch. We have shown that, on the basis of statistically viable visualizations of semantic fields, we can observe semantic differences between translated and non-translated language. The observation of these differences is potentially useful for Translation Studies, as differences on the semantic level can possibly inform us about the underlying mechanisms that cause translated language to be ultimately different from non-translated language.

On the methodological level, it is clear that further research is needed to verify the stability of the visualizations presented in this study before we can focus on a

more fine-tuned interpretation of the semantic fields. A number of the alternative methodological possibilities will need to be tested before we can proceed to a deeper level of analysis of the semantic fields. For example, a comparison of our results (based on the translational hypothesis) with results for the same data based on a distributional hypothesis (which relies on context words) could serve as a useful assessment of the stability of this translational method and could be seen as a first step towards a more fixed visualization method for semantic research in translation. A first comparison carried out to that effect by Vandevoorde et al. (2016) showed that the distributional and the translational method yield similar visualizations of the semantic field of inchoativity in Dutch. It seems, however, plausible to further use the proposed methodology to test (some of) the alleged 'universals' of translation on the semantic level. Based on the above-described exploration of the semantic fields of inchoativity in translated and non-translated Dutch, we can already formulate some hypotheses concerning the presence of the so-called 'universal tendencies' of translation on the semantic level.

Based on the above comparison of the semantic field of SourceDutch inchoativity to the fields of TransDutch, our first hypothesis about the impact of translation on the meaning distinctions of a word runs as follows: we will expect the semantic field of a lexeme under study to have fewer meaning distinctions implied in translated language compared to non-translated language. In other words, we will expect a form of "levelling out" (Baker 1996, 184) to take place *on the semantic level*. According to Baker, "we can expect less variation among individual texts in a translation corpus than among those in a corpus of original texts" (177). Transposed to the idea of semantic variation, we hence expect to detect less semantic variation (fewer meaning distinctions) in translated texts compared to original, non-translated texts.

Second, we expect to observe a specific source-language influence on the organization of the lexemes in the semantic fields of translated language, a phenomenon that would correspond to semantic "shining through." According to Teich (2003, 61–62), "translations [can differ] from comparable texts in the same language because the source language shines through." From the above description of the field of inchoativity, we could tentatively interpret the separate clustering of *begin* and *opstarten* in $\text{TransDutch}_{\text{ENG}}$ as "shining through." In order to thoroughly investigate source-language influence, additional visualizations would, however, be necessary, with the source- and target-language lexemes visualized in the same dendrogram so as to enable an inspection of the specific source-language influence on the organization of the lexemes in translated language.

A third and perhaps at first sight bolder hypothesis is that we expect the semantic representations of translated semantic fields of languages that are lexically more similar to the source language (e.g., translated English) to deviate more

from the original, authentic language (e.g., SourceDutch) than the representations of translated semantic fields of languages that are further away (e.g., translated French) from the original, authentic language. Based on the description of the semantic field of inchoativity, it was suggested that the impact of close cognate translations might affect the translated semantic field in such a way that it deviates more from the source-language field when source and target language are more similar ($\text{TransDutch}_{\text{ENG}}$ versus $\text{TransDutch}_{\text{FR}}$). Although more dissimilar languages would intuitively be thought to be more subject to alterations in semantic structure, the above hypothesis can be linked to Paradis' idea of "direct transcoding." According to Paradis' "neurolinguistic theory of bilingualism" (2004), translation will either take place "via the conceptual system" or via "direct transcoding." When a translator carries out a translation task, two scenarios are then imaginable. First, the translator's mind can function from the source-language subsystem and arrive, via the common (non-linguistic) conceptual system, to select a translation in the target-language subsystem. This "strategy" is called "translating via the conceptual system" (House 2013, 54–55; 2015; 2016, 119–120). In the second scenario, due to the considerable formal and/or semantic overlap between the source-language word and a given target-language word (a cognate), a word is activated simultaneously in the source-language subsystem as well as its cognate in the target-language subsystem. Hence, the translator arrives directly from the source-language subsystem to the target-language subsystem without processing via the common conceptual system. This second "strategy" is called "direct transcoding" (House 2013, 54–55; 2015, 119–120). The existence of these two strategies, and especially of the second one, could then explain why the semantic field of inchoativity in $\text{TransDutch}_{\text{ENG}}$ seems to deviate more from SourceDutch than the $\text{TransDutch}_{\text{FR}}$ semantic field: the more close-cognate pairs there exist between two languages for a specific semantic field, the more likely "direct transcoding" will take place and the more likely this will influence the structure of the translated semantic field.

To conclude, we hope to have shown that an explorative study such as the one presented in this paper can open the way for more research into semantic variation in translation. The above-mentioned hypotheses could be useful starting points for corpus-based or experimental studies into semantic differences in translation, allowing for a further, more in-depth exploration as well as a better understanding of the possible influence translation can have on the semantic level.

Acknowledgements

We would like to thank the two anonymous reviewers for their insightful critical comments on an earlier version of this paper.

References

Aijmer, Karin, and Anne-Marie Simon-Vandenbergen. 2004. "A Model and a Methodology for the Study of Pragmatic Markers: The Semantic Field of Expectation." *Journal of Pragmatics* 36 (10): 1781–1806. doi: 10.1016/j.pragma.2004.05.005

Altenberg, Bengt, and Sylviane Granger. 2002. "Recent Trends in Cross-Linguistic Lexical Studies." In *Lexis in Contrast: Corpus-Based Approaches*, ed. by Bengt Altenberg and Sylviane Granger, 3–48. Amsterdam: John Benjamins. doi: 10.1075/scl.7.04alt

Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511801686

Baker, Mona. 1996. "Corpus-Based Translation Studies: The Challenges That Lie Ahead." In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, ed. by Harold Somers, 175–186. Amsterdam: John Benjamins. doi: 10.1075/btl.18.17bak

Bednarek, Monika. 2008. "Semantic Preference and Semantic Prosody Re-examined." *Corpus Linguistics and Linguistic Theory* 4 (2): 119–139. doi: 10.1515/CLLT.2008.006

Benjamin, Walter. (1923) 2004. "The Task of the Translator." Translated by Harry Zohn. In *The Translation Studies Reader*. Second edition, ed. By Lawrence Venuti, 75–85. London: Routledge.

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Dam-Jensen, Helle, and Karin Korning Zethsen. 2008. "Translator Awareness of Semantic Prosody." *Target* 20 (2): 203–221. doi: 10.1075/target.20.2.02dam

Delaere, Isabelle, Gert De Sutter, and Koen Plevoets. 2012. "Is Translated Language More Standardized Than Non-translated Language? Using Profile-Based Correspondence Analysis for Measuring Linguistic Distances between Language Varieties." *Target* 24 (2): 203–224. doi: 10.1075/target.24.2.01del

De Sutter, Gert, Marie-Aude Lefer, and Isabelle Delaere, eds. 2017. *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: Walter de Gruyter. doi: 10.1515/9783110459586

Divjak, Dagmar, and Nick Fieller. 2014. "Cluster Analysis. Finding Structure in Linguistic Data." In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, ed. by Dylan Glynn and Justyna Robinson, 405–441. Amsterdam: John Benjamins. doi: 10.1075/hcp.43.16div

Divjak, Dagmar, and Stefan Gries. 2009. "Corpus-Based Cognitive Semantics. A Contrastive Study of Phasal Verbs in English and Russian." In *Studies in Cognitive Corpus Linguistics*, ed. by Barbara Lewandowska-Tomasczyk and Katarzyna Dziwirek, 273–296. Frankfurt am Main: Peter Lang.

Diwersy, Sascha, Stefan Evert, and Stella Neumann. 2014. "A Weakly Supervised Multivariate Approach to the Study of Language Variation." In *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*, ed. by Benedikt Szmrecsanyi and Bernhard Wälchli, 174–204. Berlin: De Gruyter. doi: 10.1515/9783110317558.174

Dyvik, Helge. 1998. "A Translational Basis for Semantics." In *Corpora and Cross-Linguistic Research: Theory, Method, and Case Studies*, ed. by Stig Johansson and Signe Oksefjell, 51–86. Amsterdam: Rodopi.

Dyvik, Helge. 2004. "Translations as Semantic Mirrors. From Parallel Corpus to Wordnet." In *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, ed. Karin Aijmer and Bengt Altenberg, 309–326. Amsterdam: Rodopi.

Dyvik, Helge. 2005. "Translations as a Semantic Knowledge Source." In *The Second Baltic Conference on Human Language Technologies. Proceedings*, ed. by Margit Langemets and Priit Penjam, 27–38. Tallinn: Institute of Cybernetics (Tallinn University of Technology), Institute of the Estonian Language.

Greenacre, Michael. 2007. *Correspondence Analysis in Practice*. Second edition. Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9781420011234

Greenacre, Michael. 2013. "Hierarchical Cluster Analysis." In *Multivariate Analysis of Ecological Data*, ed. by Michael Greenacre and Raul Primicerio, 89–98. Bilbao: Fundación BBVA.

Gries, Stefan. 2012. "Behavioral Profiles: A Fine-Grained and Quantitative Approach in Corpus-Based Lexical Semantics." In *Methodological and Analytic Frontiers in Lexical Research*, ed. by Gonia Libben, Gary Jarema and Chris Westbury, vol. 47, 57–80. Amsterdam: John Benjamins. doi: 10.1075/bct.47.04gri

Gries, Stefan. 2013. *Statistics for Linguistics with R. A Practical Introduction*. Second edition. Berlin: De Gruyter Mouton. doi: 10.1515/9783110307474

Haeseryn, Walter, et al. 2012. *Algemene Nederlandse Spraakkunst*. http://ans.ruhosting.nl/.

House, Juliane. 2013. "Towards a New Linguistic-Cognitive Orientation in Translation Studies." *Target* 25 (1): 46–59. doi: 10.1075/target.25.1.05hou

House, Juliane. 2015. *Translation Quality Assessment. Past and Present*. London: Routledge.

House, Juliane. 2016. *Translation as Communication across Languages and Cultures*. New York: Routledge.

Jenset, Gard B., and Barbara McGillivray. 2012. "Multivariate Analyses of Affix Productivity in Translated English." In Oakes and Ji 2012, 301–324.

Kenny, Dorothy. 1998. "Creatures of Habit? What Translators Usually Do with Words." *Meta* 43 (4): 515–523. doi: 10.7202/003302ar

Korning Zethsen, Karin. 2008. "Corpus-Based Cognitive Semantics: Extended Units of Meaning and their Implications for Translation Studies." *Linguistica Antverpiensia. New Series. Themes in Translation Studies* 7: 249–262.

Lebart, Ludovic, André Salem, and Lisette Berry. 1998. *Exploring Textual Data*. Dordrecht: Kluwer Academic Publishers. doi: 10.1007/978-94-017-1525-6

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. "Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus." *Meta* 56 (2): 374–390. doi: 10.7202/1006182ar

Mortier, Liesbeth, and Liesbeth Degand. 2009. "Adversative Discourse Markers in Contrast: The Need for a Combined Corpus Approach." *International Journal of Corpus Linguistics* 14 (3): 338–366. doi: 10.1075/ijcl.14.3.03mor

Oakes, Michael, and Meng Ji, eds. 2012. *Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins. doi: 10.1075/scl.51

Oostdijk, Nelleke, Martin Reynaert, Veronique Hoste, and Ineke Schuurman. 2013. "The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch." In *Essential Speech and Language Technology for Dutch: Results by the Stevin-Programme*, ed. by Peter Spyns and Jan Odijk, 219–247. Berlin: Springer. doi: 10.1007/978-3-642-30910-6_13

Paradis, Michel. 2004. *A Neurolinguistic Theory of Bilingualism* [Studies in Bilingualism, 18]. Amsterdam: John Benjamins. doi: 10.1075/sibil.18

Plevoets, Koen. 2015. "Svs: Tools for Semantic Vector Spaces." Ghent University.

Quirk, Randolph, Sidney Greenbaum, Geoffrey N. Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of English Language*. London: Longman.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schmid, Hans-Jörg. 1993. *Cottage und co., Idea, Start vs. Begin*. Tübingen: Max Niemeyer. doi: 10.1515/9783111355771

Schmid, Hans-Jörg. 1996. "Introspection and Computer Corpora. The Meaning and Complementation of Start and Begin." In *Proceedings of the Seventh Symposium on Lexicography*, ed. by Arne Zettersten and Viggo Hjornager Pedersen, 223–239. Tübingen: Max Niemeyer.

Simon-Vandenbergen, Anne-Marie. 2013. "English Adverbs of Essence and Their Equivalents in Dutch and French." In *Advances in Corpus-Based Contrastive Linguistics: Studies in Honour of Stig Johansson*, ed. by Karin Aijmer and Bengt Altenberg, 83–102. Amsterdam: John Benjamins. doi: 10.1075/scl.54.06sim

Simon-Vandenbergen, Anne-Marie, and Karin Aijmer. 2007. *The Semantic Field of Modal Certainty: a Corpus-Based Study of English Adverbs*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110198928

Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. "Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering." *Bioinformatics* 22 (12): 1540–1542. doi: 10.1093/bioinformatics/btl117

Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton De Gruyter. doi: 10.1515/9783110896541

Vandevoorde, Lore. 2016. *On Semantic Differences. A Multivariate Corpus-Based Study of the Semantic Field of Inchoativity in Translated and Non-translated Dutch*. PhD thesis Ghent University.

Vandevoorde, Lore, Pauline De Baets, Els Lefever, Koen Plevoets, and Gert De Sutter. 2016. "Distributional and Translational Solutions to the Visualization of Semantic Differences between Translated and Non-translated Dutch." In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*.

Vinay, Jean-Paul, and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais*. Paris: Didier.

Vossen, Piek, Isa Maks, Roxane Segers, Hennie van der Vliet, and Hetty van Zutphen. 2008. "The Cornetto Database: Architecture and Alignment Issues of Combining Lexical Units, Synsets and an Ontology." Paper presented at the Fourth International GlobalWordNet Conference, Szeged, Hungary.

Xiao, Richard, and Tony McEnery. 2006. "Collocation, Semantic Prosody, and Near Synonymy. A Cross-Linguistic Perspective." *Applied Linguistics* 27 (1): 103–129. doi: 10.1093/applin/ami045

Authors' addresses

Lore Vandevoorde
Ghent University
Department of Translation, Interpreting and Communication
Groot-Brittanniëlaan 45
9000 GHENT
Belgium

Lore.Vandevoorde@UGent.be
https://orcid.org/0000-0003-4593-4467

Koen Plevoets
Ghent University
Department of Translation, Interpreting and Communication
Groot-Brittanniëlaan 45
9000 GHENT
Belgium

Koen.Plevoets@UGent.be

Els Lefever
Ghent University
Department of Translation, Interpreting and Communication
Groot-Brittanniëlaan 45
9000 GHENT
Belgium

Els.Lefever@UGent.be

Gert De Sutter
Ghent University
Department of Translation, Interpreting and Communication
Groot-Brittanniëlaan 45
9000 GHENT
Belgium

Gert.DeSutter@UGent.be

Publication history

Date received: 22 December 2014
Date accepted: 26 July 2017
Date published: 20 September 2017