Promotor	Prof. dr. Gert De Sutter
	Vakgroep Vertalen, tolken en communicatie
Copromotor	Dr. Koen Plevoets
	Vakgroep Vertalen, tolken en communicatie

Decaan	Prof. dr. Marc Boone
Rector	Prof. dr. Anne De Paepe

UNIVERSITEIT
GENT

Faculteit Letteren & Wijsbegeerte

Lore Vandevoorde

On semantic differences

A multivariate corpus-based study of the semantic field of inchoativity in translated and non-translated Dutch

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de vertaalwetenschap

2016

Acknowledgements

On ne voit bien qu'avec le cœur
L'essentiel est invisible pour les yeux
(Antoine de Saint-Exupéry)

If I could fetch the words swirling around in my head right now in a word cloud, gratitude would be at the very center. More than just a word, it's an overwhelming feeling that fulfills me, way beyond what words can express. But being the poor human I am, with only this empty page at my disposal, let me make an attempt to express my feelings of gratitude to those without whom I could not have written this dissertation.

First of all, my sincerest gratitude goes to my supervisor, Gert De Sutter. Thank you for giving me the time and the space to explore the research field I had stumbeled into in such a surreal way. Thank you for your relentless enthousiasm, your ever pertinent remarks, your time – despite your crazy schedule and work load – but above all, your unconditional support.

To my co-supervisor, Koen Plevoets, I am much indebted to you for the development of the statistical part of this dissertation. Our long and sometimes very lively discussions where my non-stats brain met your stats brain were crucial for the development of the method which is at the core of this dissertation.

I have been blessed with the most fabulous doctoral advisory committee one could wish for. Thank you Sandra Halverson, Dagmar Divjak and Els Lefever. Your expertise has been of inestimable value to me, your remarks and insights have guided me throughout this project.

I would like to thank my colleagues of B1.16. Thank you Isabelle, for your pragmatic solutions to virtually any office-related problem (and beyond!); you are much missed (and your solutions too). To Annelore, for being there, only a chair twist away, for the numerous talks, for sharing with me. To Lynn, for being the first person in our office ever keeping a plant alive (and flourishing!), for making Mondays sweeter with chocolate cookies from Jeroen (who else?). To Pauline, for giving me that warm feeling of being connected to someone. Thank you for the many laughs, your clumsiness beyond what even I thought was possible; for your perseverance too, which makes it so great working with you. Thank you to Chloé, for offering your help whenever you could, especially taking books back up the Blandijnberg.

I am grateful to a number of other colleagues too, who've made me feel welcome and at home in the *talenhuis* I had left behind me six years before. Thank you Stefaan and the other members of the *afdeling Nederlands*. Thank you Christophe, Kristien and Alexandra, for getting me through the first difficult months. To Sofie for rescue chocolate, to Sevdag for not getting me off the AWD list despite my many absences. Thank you Katrien, for keeping an eye on my finances; Valérie, for emergency scans and finding back the most hidden works of the library; Bert, for your kind help and your efforts to order new books at the speed of academic lighting. To Peter, for saving my computer from fruit tea and me from a nervous breakdown.

A special *merci* to Désirée. You were the first person to believe in me and although this PhD is lightyears from what we've imagined I would work on, I hope you feel as I do today, *un tout petit peu fière quand même*. Thank you to Hildegard and June, for hiring me in the first place and trying to find the best solution when things went wrong before they even started.

To my friends, the new ones I made, the old ones who stayed, thank you. Ann, for being the best train buddy ever (you're missing out on first class travelling) and becoming such a thoughtfull friend. Irene, for your warmth and sincere friendship, for homemade pasta and wedding stories. Carol, for cava *apero's in* your sunny garden in summer or in the coziest kitchen of Ghent in winter, the most beautiful pretext of friendship refound. Elisa, for teaching me gratitude and always being so close; distance is no barrier for the heart. Christina, for long talks, long walks (and bikerides!) and inspiring me with your view on life. Jasper, for sincerely disencouraging me to step into this – it didn't work, but at least I was warned. To Sylvie, Bjoke, Evy, An-Sofie, Camille, Ruben, Chen, Lieve, Jan, Annelies, Nico, Hannelore, Jan, Lien and Kevin for caring about me, for being there in moments much more complex than a PhD writer's block. To Valérie, *merci*, for your little phonecalls and words of understanding. To Anne et Francis, for the best *galettes* ever, brightening up my coffee breaks during this last difficult month.

I want to thank my parents, who gave me the precious freedom to study whatever I wanted, for as long as I wanted to (little did they know): you thought me to work hard, to *try harder* and I did. Mama, we've beaten the odds. Papa, you're the best one-man rescue team on the planet. Jasmien, thank you for always caring, for birthday presents on my desk and being the most creative, sensible and perseverant person I know (and thank you Kenny, for looking after her).

Pour Fabrice, parce que mon cœur chante en français pour toi, et que je ne pourrais autrement te dire merci que dans cette langue que nous partageons. Merci d'avoir été et de toujours être là, à mes côtés.

List of Abbreviations

CA	Correspondence Analysis
CBTS	Corpus-based Translation Studies
CL	Corpus Linguistics
DPC	Dutch Parallel Corpus
DTS	Descriptive Translation Studies
HAC	Hierarchical Agglomerative Clustering
HCA	Hierarchical Cluster Analysis
LPTs	Linguistically Predictable Translations
MC	Mutual Correspondence
MCA	Multiple Correspondence Analysis
NLP	Natural Language Processing
SMM	Semantic Mirrors Method
SMM++	Extended Semantic Mirrors Method
TS	Translation Studies
WSD	Word Sense Disambiguation

List of Tables

List of Figures

Table of Contents

Chapter 1
Introduction

Within the discipline of corpus-based translation studies, it is often assumed that translated texts incorporate typical linguistic features which distinguish them from non-translated, original texts (Baker 1993). These typical linguistic features – also called *translation universals* – are defined as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker 1993, 243). Over the last two decades, numerous corpus-based studies have either validated or refuted one or more of these universals (Malmkjaer 1997; Laviosa 1998, 2002; Mauranen 2000; Olohan & Baker 2000; Baker 2004; Bernardini & Ferraresi 2011; Delaere et al. 2012; De Sutter et al. 2012 – see Kruger 2012 for an overview). The bulk of these studies have focused on lexical and grammatical phenomena, but translational features on the *semantic level* have been somewhat neglected (Laviosa 2002, 28)[1] within translation studies. On the one hand, this 'lacuna' is quite remarkable in the sense that the notion of meaning has always been at the core of translation as a task as well as of translation studies as a discipline (see for example the *Translation and Meaning* series, discussed in section 2.2.3.1). On the other hand, the exclusion of the semantic level from the universals debate does not come as a total surprise. For one thing, semantic differences are difficult to operationalize 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 TS, dismissing it at once as a subject of research (see section 2.2.2.6). However, the identification of semantic differences between translated and non-

[1] This does not mean that the role of semantics itself in translation has not been addressed (e.g., Klaudy, 2010), but this kind of research is rarely corpus-based and barely ever involves with denotational issues.

translated language offers the prospect of revealing hidden cognitive processes which could, in their turn, possibly explain these alleged *universals* of translation on the cognitive level (and offer an alternative to the hitherto suggested explanations of diverging norms and conventions).

Hence, the goal of this dissertation is to formulate a first, tentative answer to the question whether (some of) the universal tendencies of translation (we will focus on *levelling out*, *normalization* and *shining through*) also exist on the semantic level. By comparing visual representations of semantic fields in different varieties of Dutch (non-translated Dutch, Dutch translated from English and Dutch translated from French), we want to gain a better insight into the impact translation possibly has on the structure of those semantic fields (and whether this impact can be understood in terms of *universal* translation tendencies). To do so, we will present a quantitative bottom-up corpus-based method which will enable us to measure and to visualize semantic similarity within semantic fields representing translated language and others representing non-translated language. The proposed method (the Extended Semantic Mirrors Method) builds on an existing method for automatic thesaurus building, the Semantic Mirrors Method (Dyvik 1998; 2004; 2005) which has been successfully implemented within contrastive linguistics to discern semantic fields (Aijmer & Simon-Vandenbergen 2004, 2006; Simon-Vandenbergen 2013). Our extended method proposes an accurate, statistics-based visualization of the observed fields and is furthermore applicable to research in TS.

1.1 The semantic field of *beginnen* / inchoativity

To date, very little research has been conducted on semantic differences in translation, making it difficult to formulate clear hypotheses. This explains the explorative nature of our study, which will make use of statistical visualization tools that are specifically designed for such kind of data exploration. The possibilities to investigate semantic differences via corpora are potentially limitless. For this study, we choose to investigate the semantic field of *beginnen* [to begin] / inchoativity. Our aim is then to reveal possible semantic differences between the semantic field of *beginnen* / inchoativity in non-translated Dutch and two semantic fields of translated Dutch; translated from English and translated from French respectively (in accordance with the available languages in the Dutch Parallel Corpus which will be used for this study). The choice of 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 viz. *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 e.g. 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.

1.2 Research questions

The question whether the so-called universal features of translation also apply to the level of meaning has to our knowledge rarely or never been posited. As we will see in sections 2.2.2.3 and 2.2.2.4, *levelling out*, *normalization* and *shining through* seem to be the best candidates to investigate the differences and similarities of the semantic relationships in (general) translated Dutch (we will investigate translated Dutch with English as source language – TransDutch$_{ENG}$ – and with French as a source language – TransDutch$_{FR}$) compared to (general) non-translated Dutch (SourceDutch). Before we can formulate our research questions with respect to the *universals* of translation on a semantic level, we need to distinguish between two possible perspectives of study in semantic research.

In lexical semantics, a distinction is usually made between studies which take a semasiological outlook and others which take an onomasiological outlook on meaning (Geeraerts et al. 1994). Semasiology takes the point of view of the different concepts which can be expressed by one word (the polysemy of a word); onomasiology takes the viewpoint of the different words that can be employed to express a single concept (near-synonymy). Given our choice to conduct this study on the most prototypical expression of inchoativity in Dutch, *beginnen*, both a semasiological and an onomasiological outlook are possible.

A semasiological outlook implies that the intended visualizations are considered as possible and plausible representations of the different meanings of a word under study (in our case *beginnen*). In this case, the 'representation of different meanings' of a word are considered as a semantic map, "a representation of meanings or uses and the relations between them" (Simon-Vandenbergen & Aijmer 2007, 23, following van der Auwera & Plugian). From an onomasiological point of view, our visualizations would then represent the different ways of expressing one and the same concept under study (in our case, the field of *inchoativity*).

If we want to discover which are the different words that can be used to express the concept of inchoativity (onomasiological viewpoint), the best option, in a corpus study such as ours which typically does not give direct access to concepts but (only) to words i.e. to lexicalizations of those concepts, would be to start with its most prototypical expression. On the other hand, the fact that our study starts off with a single word, i.e. *beginnen*, simultaneously favors a semasiological outlook on meaning. If we want to explore the different concepts expressed by *beginnen*, the most logical choice would be to start our study with this lexeme itself. Hence, the choice of the initial lexeme *beginnen* allows us to take both a semasiological and an onomasiological outlook. We do acknowledge the necessity of distinguishing the two perspectives, although they are closely interwoven. Geeraerts (2010, 30) reminds us that "the semasiological extension of the range of meanings of an existing word is itself one of the major mechanisms of onomasiological change – one of the mechanisms, that is, through which a concept to be expressed gets linked to a lexical expression". Therefore, the link between a lexical expression and a concept is always semasiological in one direction (from lexical expression to the (range of) concept(s)) and onomasiological in the other direction (from the concept to the (range of) lexical expressions).

As we will see in section 3.6.2, our visualizations will correspond to the visual output of a statistical analysis via Hierarchical Agglomerative Clustering. We will consider the different groupings (clusters) in a visual representation (dendrogram) as different meaning distinctions of the word under study. In particular, this means that each cluster in the dendrogram will be considered as a separate meaning (a meaning distinction) of the semantic field of the word under study (*beginnen*) (semasiological outlook). In addition, we will consider the lexical items which make up each cluster as the lexical expressions of the particular meaning distinction of the cluster they belong to (onomasiological viewpoint). It is also possible to take a broad onomasiological outlook and to consider each visualization (dendrogram) as a whole as a representation of a semantic field of inchoativity. The lexical items in the visualizations are then considered as lexical expressions of the central concept of inchoativity. This second option would imply that somewhat less importance is given to the actual clustering: rather than considering the clusters as meaning distinctions of the central word, the clusters would 'simply' indicate which lexemes are more near-synonymous expressions of the central concept. We choose to take the double semasiological-onomasiological outlook here (clusters as meaning distinctions of the central word and lexical items in each cluster as the expressions of the meaning distinction of the cluster) because this double view can possibly allow us to comprehend whether the universal tendencies of translation are taking place on the semasiological level of the different meanings of a word (can the polysemy of a word be altered under influence of translation?), or on the onomasiological level of the words expressing a particular meaning distinction (is the near-synonymy relation between different words altered under influence of translation?). Within this outlook, it remains at all times possible to consider the semantic field as a whole as a representation of the semantic field of inchoativity, represented by its (most prototypical) means of expression. In this case, less attention is paid to the meaningfulness – in terms of meaning distinctions of a central word – of the clustering.

In the remainder of this section, we present the research questions about the presence of the universal tendencies of *levelling out*, *shining through* and *normalization* in (general) translated Dutch (TransDutch$_{ENG}$ and TransDutch$_{FR}$) compared to (general) non-translated Dutch (SourceDutch) for the semantic field of *beginnen* / inchoativity. For each question, we will also indicate how bottom-up statistical visualizations can inform us about the presence or absence of these typical translation properties within the semantic fields.

Firstly, on a semasiological level: **do the meanings expressed by *beginnen* (or does the prototype-based organization of those meanings) differ in translated language compared to non-translated language?** Does this difference consist in *beginnen* having fewer different meanings implied in translated language compared to *beginnen* in non-translated language? If this is the case, we can call the phenomenon semasiological *levelling out*.

Semasiological levelling out can be investigated by comparing the variation of a certain feature in translated language to the variation of the same feature in non-translated language (see section 2.2.2.4). Given the presumed subtlety of the meaning distinctions for *beginnen*, we expect the semantic variation between the fields of translated and non-translated Dutch to be small and hence difficult to observe by mere inspection of the clusters in the dendrograms. We will therefore measure the centrality of each of the meanings and focus on possible changes within the prototype-based organization of the clusters. In order to assess changes in the prototype-based organization of the meanings within the semantic fields, we will evaluate the distance from each of the meanings (clusters) in a field to the center of the semantic space to see whether in translated language (TransDutch$_{ENG}$ and TransDutch$_{FR}$), some meanings (clusters) become more peripheral and others more central compared to non-translated language (SourceDutch).

If we indeed observe differences in the prototype-based organization of the meanings in translated and non-translated language, **could there be a) an influence of the source language (*shining through*) on the translated language or b) will the expressed meanings in translated language conform to (the organization of) the meanings expressed in non-translated language (*normalization*)?**

In order to investigate source language influence (*shining through*) on the semasiological level, meaning distinctions in translated language need to be compared to meaning distinctions present in the source language of the translation. To do so, we will visualize the semantic fields of the closest equivalents of *beginnen* in the source languages of TransDutch$_{ENG}$ and TransDutch$_{FR}$ (SourceEnglish *to begin* and SourceFrench *commencer*). We will compare the different meaning distinctions (clusters) in the fields of *to begin* and *commencer* to the meaning distinctions (clusters) in translated language (TransDutch$_{ENG}$ and TransDutch$_{FR}$) to see whether the specific (prototype-based organization of the) meaning distinctions within the semantic fields of SourceEnglish and SourceFrench have (has) influenced the organization of the meaning distinctions in translated language (TransDutch$_{ENG}$ and TransDutch$_{FR}$).

Target language influence (*normalization* effects) will be investigated on the semasiological level by comparing the prototype-based organization of the different meanings expressed by *beginnen* in a semantic field of translated language (TransDutch$_{ENG}$ *and* TransDutch$_{FR}$) to the semantic field of non-translated language (SourceDutch) to see whether the prototype-based organization of the meaning distinctions (clusters) in translated language conforms to the organization of those in non-translated language. This comparison is particularly interesting when contrasted with the observations about semasiological *shining through* since *shining through* and *normalization* are often understood as the two extremities in a continuum (Hansen-Schirra & Steiner 2012, 272) between source- and target language influence.

Secondly, on an onomasiological level: **Will the words expressing the concept of inchoativity (or the prototype-based organization of those words) differ in translated Dutch compared to non-translated Dutch?** We will focus on the possible changes in the prototype-based organization *of the lexemes within each cluster*[2] (representing a particular meaning distinction). We will assess the distance from each of the lexemes within a cluster to the center of the cluster it belongs to, to see whether in translated language (TransDutch$_{ENG}$ and TransDutch$_{FR}$), some lexemes become more peripheral and others more central within a particular meaning distinction compared to non-translated language (SourceDutch). Our method will however not allow us to investigate whether a given concept is expressed by *fewer* lexemes in translated Dutch compared to the same concept in non-translated Dutch (onomasiological *levelling out*), because the total number of lexemes within each semantic field is kept stable over all visualizations (see section 3.4.3). Observations on the onomasiological level will inform us about changes in near-synonymy relationships (based on changes in prototype-based organization)

[2] In contrast to the assessment of the *clusters* within each *dendrogram* on the semasiological level.

between the lexemes in the semantic field, but we will not directly connect these observations to onomasiological *levelling out*.

If we indeed observe differences in prototype-based organization of the lexemes within the clusters, **do we rather see a) the organization of the lexemes in the source language semantic field *shine through* in the translated semantic field; or, b) will the organization of the lexemes within the clusters (meaning distinctions) in translated language tend to be more similar (*normalize*) to the organization of the lexemes within the meaning distinctions in non-translated target language?**

We will investigate source language influence (*shining through*) on the onomasiological level by visualizing the French and English source language lexemes together with the Dutch target language lexemes. In particular, it will be measured and visualised to what extent the French and English source language lexemes determine the clustering of the Dutch lexemes into specific meaning distinctions. In this way, we can see whether the specific organization of the lexical items in the clusters – with each cluster representing a particular meaning distinction of *beginnen* – is possibly influenced by a specific underlying source language lexeme.

Target language influence on the onomasiological level (*normalization*) can again be investigated by comparing the prototype-based organization of the lexemes within each of the meaning distinctions in TransDutch$_{\text{ENG}}$ and TransDutch$_{\text{FR}}$ to that of SourceDutch. Again, this comparison is all the more interesting when contrasted with the insights about onomasiological *shining through*.

Just as for the semasiological level, we expect the changes on the onomasiological level to be rather small as well. Keeping furthermore in mind that we are presenting a first, tentative study of universal tendencies of translation on the semantic level, we do by no means pretend to reveal general tendencies of onomasiological *levelling out*, *shining through* or *normalization* as such kind of conclusions would require much more further research into ‘semantic universals of translation’.

1.3 Outline of the dissertation

Chapter 2 provides the theoretical foundations for this dissertation. Our central research question “does translated language differ from non-translated language on the semantic level?” is embedded within the discipline of corpus-based translation studies (hence: CBTS). The first part of our theoretical chapter (section 2.2) will therefore zoom in on a number of important aspects of CBTS. We will look into different sub-disciplines of linguistics in an attempt to gather a number of theoretical building blocks necessary to investigate meaning relationships in translation. We will moreover elaborate on our choice of a corpus approach and hence the need for a representative corpus. Next, we will equally discuss (i) the place of the study of meaning within the wider area of CBTS and the seemingly problematic relation between the study of *universals* and the study of meaning; (ii) the importance of a contrastive linguistic procedure such as *back-translation* and (iii) the relationship between universals and meaning and the notion of *equivalence* (a notion which we will explore in various sub-disciplines, viz. translation studies, contrastive linguistic studies as well as in computational investigations of meaning). The second part of this chapter (sections 2.3. and 2.4) will provide the theoretical foundations for the development of a bottom-up, statistical visualization method of semantic fields in both translated and non-translated language. We will zoom in on the possibilities offered by the existing technique of *semantic mirroring* which uses the procedure *back-translation*; the usefulness of statistical techniques for visualization purposes and the necessity of a theoretical framework within which the created visualizations can be interpreted.

In chapter 3, the method which will be applied in this dissertation is described thoroughly. Our method will consist of two extensions of an existing method, the Semantic Mirrors Method (Dyvik 1998; 2005). A first extension will allow us to retrieve candidate lexemes for semantic fields of inchoativity in translated and non-translated language. A second extension will enable us to visually represent the obtained data sets. To this extent, we will propose a statistical extension of the SMM which will allow us to create visual representations of a semantic field under investigation. The ultimate aim of these extensions is to enable us to compare visualizations of semantic fields of translated and non-translated Dutch to each other so as to reveal possible semantic differences between these varieties. In this chapter, we will furthermore propose to investigate *levelling out* by looking for possible changes in the prototype-based organization of the semantic fields on the basis of measures such as *centroids* and *medoids*. In order to measure specific source language influence (*shining through*), we will determine the positions of the (English or French) source language lexemes in the Dutch semantic spaces via two additional analyses: the visualization of the source language fields of inchoativity in English and French (the pivot languages of the SMM++) and Multiple Correspondence Analysis.

In chapter 4, our method is applied to the field of inchoativity in Dutch. The results are presented and described on the basis of three main visualizations, one for a semantic field of inchoativity in non-translated Dutch, one for translated Dutch with English as a source language and one for translated Dutch with French as a source language. We will focus on the differences between the semantic field of non-translated Dutch inchoativity and the fields of translated Dutch inchoativity. Our goal is to explore the semantic fields of translated and non-translated Dutch in an attempt to reveal instances of *levelling out, normalization* and *shining through* on both the onomasiological and the semasiological level.

In chapter 5, an attempt will be made to connect the obtained results to current hypotheses in corpus-based cognitive translation studies and neurolinguistics. Two cognitive explanational hypotheses will be put forward and tentatively applied to the results of our study: the Gravitational Pull Hypothesis, developed by Sandra Halverson and the Neurolinguistic Theory of Bilingualism, developed by Michel Paradis.

Chapter 6 concludes this dissertation with an overview of our main findings with regard to the differences and similarities of the semantic relationships in (general) translated Dutch compared to (general) non-translated Dutch for the semantic field of *beginnen* / inchoativity. We will also briefly reflect upon the methodological contribution this dissertation possibly makes to the empirical study of semantics in translation, especially with regard to the impact of translation on semantic representations.

Chapter 2
Theoretical considerations

2.1 Introduction

Modern corpus linguistics as we understand it today arose as from the 1960s, in the early days of the digital age. The appearance of electronic corpora in linguistics opened up the way for the development of numerous corpus-related sub-disciplines of linguistics. In the early 1990s, the use of corpora to study translational behavior was fully acknowledged within translation studies thanks to a seminal paper by Mona Baker (1993), and the sub-discipline corpus-based translation studies (hence: CBTS) was born. It is within this paradigm that our study is situated.

In the first part of this chapter (section 2.2), we will introduce the discipline of CBTS within which this dissertation is profoundly embedded. As will appear from this section, CBTS does not offer a clear-cut methodological framework to conduct a corpus-based study of meaning relationships in translation. The theoretical, methodological and descriptive footing to develop such a method will therefore be sought within other corpus-related areas of linguistics.

In section 2.3, we will investigate a number of contrastive corpus studies. We will explore the notion of *back-translation*, a procedure which relies on *translation equivalence* and is known to reveal semantic relationships. Special attention will be given to the Semantic Mirrors Method (hence: SMM), which exploits the procedure of *back-translation* and fulfills the prerequisites to validly compare meaning relationships in translated and non-translated language.

Various sub-disciplines of corpus semantics further provide useful insights for the investigation of semantic relationships in translation. In section 2.4.1, we will elaborate on the notion of *translational equivalence*. Its operationalization within Word Sense Disambiguation (hence: WSD) can be transferred to a corpus-based translational study as a solution to the operationalizability problem of *equivalence*. Corpus-based quantitative studies typically generate large amounts of data. In order to reveal the semantic information hidden in the corpus data, we want to create bottom-up, statistical visualizations of semantic fields in translated and non-translated language. In section 2.4.2, we will see that statistical visualizations of 'that what cannot be seen by the bare eye' can be a potentially good lead towards meaningful representations of meaning relationships. In section 2.4.3, we propose to combine the corpus-based quantitative visualizations with a theoretical framework from cognitive linguistics. We will propose to use the prototype model of category structure as a necessary basis for a coherent interpretation of the statistical visualizations we aim to create.

2.2 Corpus-based translation studies

In the first part of this section (section 2.2.1), we will zoom in on the different types of corpora, which constitute the main methodological tool in CBTS. In the second part of section (section 2.2.2), we will focus on how precisely this new sub-discipline arose within translation studies, by further exploring the research program set up by Baker. We will give extensive consideration to the *translation universals* paradigm in an attempt to understand why research into *universals* on the semantic level has barely had any uptake within CBTS. Admittedly, there exists research in CBTS that focuses on alternate subjects such as individual variation, translation norms and conventions or translation language change (Zanettin 2013, 21). We choose, however, to focus on the *universals* research program which has undeniably dominated the field since the 1990s. In addition, we will determine which *universals* would seem best suited for the investigation of semantic relationships in translation. In section 2.2.3, we will focus on the so-called cognitive turn in translation studies, which enabled the re-introduction of linguistic meaning into translation studies. The central notion of *equivalence* will be discussed in section 2.2.4. As we will see, both the notions of *linguistic meaning* and *equivalence* inevitably need to re-take a central position here if we are interested in studying semantic relationships in translated and non-translated language.

2.2.1 Corpora

Corpora come in so many flavors, shapes and sizes that it is virtually impossible to give an exhaustive overview of the existing corpora today (McEnery & Hardie 2012). For learner corpora only, the Center for English corpus linguistics of the Université Catholiqué de Louvain lists close to 150 different corpora (Hiligsmann, 2015). In an attempt to structure the enormous amount of corpora that are out there, several researchers have come up with corpus typologies; e.g Johansson (1998) set out a typology for cross-linguistic research, Baker (1995) and Laviosa (2002) drew up typologies from the viewpoint of CBTS, Tognini Bonelli & Sinclair (2006), Lee (2010) and many others attempted typologies for the general purpose of CL, while numerous other overviews keep on appearing in an effort to keep up with the unceasingly growing number of corpora that is out there.

Instead of undertaking a (necessarily non-exhaustive) overview of existing corpora, we opt to lay out the different *dimensions* along which a corpus can be defined (size, content and corpus languages). A better understanding of these dimensions is indispensable for the further selection of a corpus that suits our research needs, i.e. a corpus of sufficient size that can be used to conduct research into the differences and similarities of the semantic relationships in (general) translated Dutch compared to (general) non-translated Dutch for the semantic field of inchoativity. The corpus selected according to these parameters, the Dutch Parallel Corpus, is described in section 3.2.

2.2.1.1 Size

The first electronic corpus – the Brown corpus – was established in 1961 and counted a little more than one million words. Ever since, the goal seemed to be set at building ever larger corpora. It had indeed been remarked that some (more rare) linguistic phenomena could be absent from a corpus (and could consequently not be investigated) merely because the corpus was too small, so the idea that sizes mattered (a lot) was quickly assimilated. To overcome the obstacle of corpus size, the logical step was thus to (simply) build larger corpora: from a little more than 1 million words in 1961, to the appearance of the Oxford English corpus at the turn of the millennium counting over 2 billion words (Figure 1). By that time, the world wide web had started to be used as a corpus too.

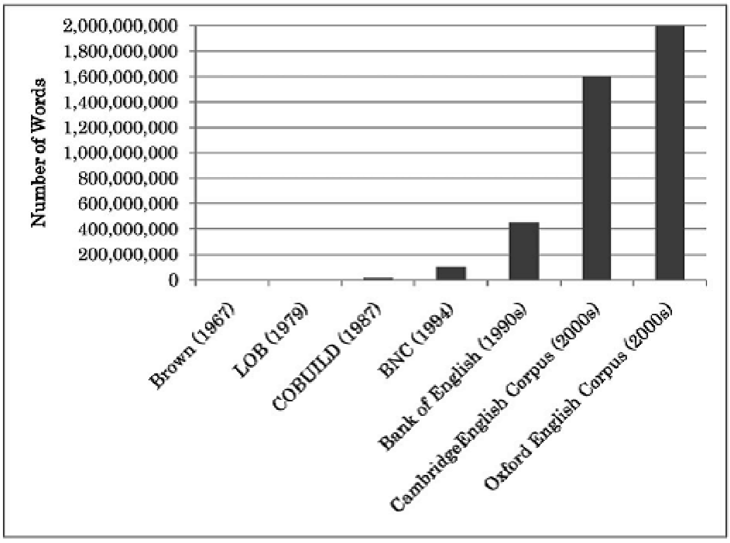


Figure 1 Growth in corpora sizes over 50 years (copied from Anthony (2013, 145)).

Over the last decades, the average size of corpora has been growing steadily, with nowadays corpora containing hundreds of millions of words. However, this trend is observed to a far lesser extent for corpora in languages other than English, and even less so for bilingual or multilingual corpora. Corpora specifically suited for the study of translation such as The English-Norwegian Parallel corpus – around 2.6 million words – (Johansson 1998), The Dutch Parallel corpus - around 10 million words – (Macken et al. 2011) or the CroCo corpus – about 1 million words – (Hansen-Schirra et al. 2012) do not generally exceed 10 million words (see also the overview by Zanettin (2013, 26-27)). Although larger corpora would have the same advantages mentioned earlier with respect to the (monolingual) English corpora – more data allow to investigate more rare linguistic phenomena that can remain unnoticed if the corpus size is too small – researchers in translation studies often have to content themselves with smaller corpora such as the ones cited above, simply because the bigger corpora that exist cannot be used for investigations in translation studies (although – usually larger – comparable corpora have been frequently used in CBTS).

2.2.1.2 Content

While for most of the history of Corpus Linguistics, definitions of what a corpus is immediately limited its content to files of text, the recent appearance of multimodal corpora (Kipp et al. 2009) has introduced other types of data-carriers such as video and (live) streaming into the corpus-world. Although this new development is uncontestably a very interesting one, we will not further explore this type of corpora (since our own study will be carried out with a corpus consisting of text files).

A great deal of dimensions with respect to the types of text files that a corpus contains, need to be defined. First, the text files can consist of written material or they can contain transcriptions of spoken language, or both. Second, the corpus can aim to be representative of general language; alternatively, it can contain different text types (the corpus can be balanced with respect to the different text types – or not), or it can be a specialized corpus, focusing on one particular text type (e.g. a corpus of legal texts). Thirdly, the corpus can be built up by complete texts or samples of texts (n words from the nth to the nth word of each text). The advantage of sampling is that "the number of words from each text can be exactly matched", making it easier for the corpus designer to arrive at equal proportions per text type (Deignan 2005, 77). The danger with sampling is that some linguistic phenomena that tend to appear at the beginning or ending of texts might not be present in a corpus built up by samples (Deignan 2005, 77 referring to Stubbs 1996). A corpus can also be a mix of samples and full texts, of course. The fourth dimension concerns the dynamic (open) / static (closed) nature of a corpus: a closed corpus is delivered as a finite product, to which no texts are further added. A dynamic, open corpus on the other hand – also called a monitor corpus –is not so finite in the sense that materials can be added over time (McEnery & Hardie 2012, 6). Both open and closed corpora can be employed for diachronic studies (of change over time) or synchronic studies (focusing on a particular period), all depending on how the corpus is used by the researcher (Johansson 1998, 3).

2.2.1.3 Language(s) of the corpus

The final dimension concerns the number of languages present in a corpus. If there is only one language represented, the corpus is a monolingual one, with two languages, it is called bilingual, and with more than two languages present in the corpus a multilingual corpus. Laviosa (2002, 36-38) has proposed a further subdivision of these three types. Her corpus typology is particularly focused on the applicability of corpora to the study of translation. Since this is also the type of research we will pursue, we will maintain her focus – bearing in mind that the dichotomy translated – non-translated could in fact be replaced by any two language varieties the researcher would wish to compare with each other:

- A monolingual corpus can be a *single monolingual corpus*, consisting of one set of texts (either translated texts or non-translated texts), in one language, whereas a *comparable monolingual corpus* consists of two monolingual corpora, one with translated and the other one with non-translated texts (all other design criteria are stable).
- A bilingual corpus can be a *comparable bilingual corpus*, consisting of two monolingual corpora in two different languages – all other design criteria are or

should be (as) stable (as possible)[3] – that can consequently be compared to each other. A *parallel bilingual corpus* then consists of texts in two different languages, with the texts in one language being the originals of the translations in the other language. *Parallel bilingual corpora* can further be mono- or bi-directional. Mono-directionality means that language A is always the source language and language B always the target language; bi-directionality implies that language A and language B can both be source and target language.

- A *comparable multilingual corpus* is similar to a *comparable bilingual corpus*, but with more than two languages involved; a *parallel multilingual corpus* is similar to a *parallel bilingual corpus*, again with the only difference of the number of languages involved. Laviosa indicates a supplementary difficulty here: *parallel multilingual corpora* can be *mono-source* – only one of the several languages is the source language, the other languages are target languages; *bi-source* – two of the several languages can be the source language; or *multi-source* – several or all of the languages in the corpus can serve as source language.

As stated above, Laviosa established her corpus typology because she considered it to be "an essential step towards developing a coherent methodology in corpus-based translation studies" (Laviosa 2002, 38). In section 3.2 we will use the terminology proposed by Laviosa to define the corpus that we will use for this study, i.e. the Dutch Parallel Corpus.

2.2.1.4 General issues with corpora

The use of corpora in linguistics – although widespread and well-accepted in present-day linguistics – does also raise a number of issues. One of the most common discussions in CL was initiated by Tognini-Bonelli (2001) and is concerned with the difference between *corpus-based* and *corpus-driven* research. Put shortly, corpus-based approaches consider corpora as a method of research, whereas corpus-driven approaches see corpora as the impetus for theoretical development in linguistics (for discussions on this topic, see Hardie & McEnery 2010, 384-385; McEnery and Hardie 2012, 150 ff.). The importance of this distinction has been questioned by Xiao (2009, 994), who finds the "sharp distinction" between corpus-based and corpus-driven approaches "overstated" and Gries (2010, 328), who argues that he sees no reason to consider CL as a theory; in his view CL is a methodological paradigm.

[3] "in an attempt to ensure that their linguistic differences can be reliably attributed to their status as translation versus non-translation (in the monolingual comparable) or to their languages (in the bi/multilingual comparable), rather than to confounding variables" (Laviosa 2002, 39).

A second issue involves *representativeness*, which is one of the most cited conditions imposed upon a corpus. This representative function can stretch from standard varieties of a language "to any kind of specialized language (represented in a domain-specific corpus)" (Leech 1991, 11). However, no corpus – irrespective of how careful the compilation process has been carried out – can ever claim *absolute* representativeness. For instance, corpora that do not explicitly claim text-genre balancedness are sometimes only representative of the journalistic text type, because this is the text type that is most easily available. Even for an (explicitly) text-type balanced corpus, we can never be sure whose language the corpus is representative of. As Deignan puts it clearly:

> Because there is such a wide variation in the range and relative proportions of text types that we each see and hear, no corpus could ever represent anyone's personal experience of language more than fleetingly. This does not have to be seen as a disadvantage; it can be argued that a well-balanced corpus is superior to an individual's personal corpus in its range and balance (Deignan 2005, 91).

The importance of *representativeness* also amounts with the type of research one wishes to conduct: it is important for a semanticist looking for the many meanings of, for instance, the lexeme *translation* to have a corpus at one's disposal that is representative of different text types so as to detect the different (metaphorical) meanings this lexeme is likely to have in different genres. Overall, if we let go of the illusive idea of *absolute representativeness*, and provided we compile / select our corpus with caution, then, a corpus built in a balanced way with respect to different text types and compiled of texts selected from a wide range of different sources can be held as the current best possible representation of a standard variety of a language.

Finally, a third issue focuses on the advantages and disadvantages of parallel corpora. Whereas parallel corpora consist of source texts and their translations, the texts in a comparable corpus are simply *comparable* to each other according to a number of parameters set by the corpus designer (e.g. text length, genre, etc.) but they are not each other's translational counterparts. The issue of comparability is the weak point of comparable corpora since "[s]ome types of text are culture-specific and simply have no exact equivalent in other languages" (Granger 2003, 19). On a micro-textual level, when using a comparable corpus, it may be difficult to know which forms in the compared languages have similar meanings and pragmatic functions, and which forms can consequently be compared and which ones not (Johansson 1998, 5). On the other hand, comparable corpora seem to be easier and faster to compile than parallel corpora since explicit identification of texts as original texts vs. translation is more pervasive than encountering a source text delivered together with its translation. In this dissertation, preference is given to the use of an existing parallel corpus since the investigation of semantic relationships requires us to be able to make comparisons on the word- level. The only drawback is that all texts labeled as original (non-translated) language in a parallel corpus have at some point been selected to be translated (since all non-translated texts in a parallel corpus are the source language text of a translated text in the corpus). This does not alter anything to the 'originality' of the original language of course, but it can have influenced the presence and/or absence of certain texts on the basis of their 'suitability' to be translated or not. In order to overcome this problem, it is possible to include a monolingual reference corpus for supplementary comparison, but studies that did so have faced major comparability issues due to corpus size or the uncertainty about the (translational) status of the texts in the presumed original language corpora (see e.g.: Förster Hegrenaes (2014)).

2.2.2 Baker's universals

The paper that has literally catapulted translation studies into the era of corpus research – although preceded by work by Toury (1980), Frawley's idea of *third code* (1984) as well as studies such as Gellerstam (1986) – was without a doubt Mona Baker's 1993 seminal article "Corpus Linguistics and Translation Studies". Baker indeed foresaw that:

> the techniques and methodology developed in the field of corpus linguistics will have a direct impact on the emerging discipline of translation studies, particularly with respect to its theoretical and descriptive branches (Baker 1993, 233).

The article provoked a true corpus turn in translation studies leading to the development of a research program that was mainly constructed on the basis of the idea of *translation universals*, equally proposed in that same article. But why was this corpus turn so much-needed in translation studies? The main reason was probably that the positing of this new paradigm within TS allowed for an emancipation of the discipline with respect to other adjacent linguistic disciplines and especially with respect to contrastive linguistics, where translations were seen as a useful methodological tool rather than an object of study (see section 2.3). Baker assigns a new and prominent role to parallel and in particular to comparable corpora: instead of dismissing translations as "second-hand and distorted versions of 'real' texts" (Baker 1993, 233), she puts them at the center of attention, claiming that the interest for TS is precisely to study in what way translations, as "genuine communicative events and as such [...] neither inferior nor superior to other communicative events in any language" (Ibid., 234) differ from non-translations. She asserts that a number of preparatory parameters needed to be set (e.g. the introduction of corpora in TS) so that this type of research could actually come into being:

> There is now an urgent need to explore the potential for using large computerized corpora in translation studies. It seems to me that most of the components for

> realizing this potential are in place. The emphasis has shifted *from meaning to usage*, and the notion of *equivalence is gradually giving way to that of norms*. The *status of the source text* has been undermined and we have managed to make the leap from source-text-bound rules and imperatives to descriptive categories. There is increasing interest *in features of translated texts* per se and we are beginning to develop *a descriptive branch* of the discipline with well-defined objectives and an explicit program. [...] A suitable *methodology* and a set of very powerful and adaptable tools are now available from corpus linguistics (Baker 1993, 248, all emphases are ours).

Baker urges researchers to move over from a *prescriptive* to a *descriptive* branch of TS and to do so via the methodology and tools of corpus linguistics. Instead of proposing or imposing rules on how one should translate or to prescribe what translation *should be*, TS needs to explore what translation *is* by investigating the actual *usage* in translation and by exploring the specific *features* of translated texts. In this respect, Baker sees the need of dismissing terms such as *equivalence*, *correspondence* and *shifts* "which betray a preoccupation with practical issues such as the training of translators" (Baker 1993, 235). The fact that she actually *can* dismiss those terms has to do with another proposed attention shift : instead of focusing on the source text – which in Baker's view is precisely the *source* of the rule-governedness and prescriptive nature of TS – she proposes to focus on the target text, i.e. the translated texts themselves and their features. The dismissal of the terms *equivalence*, *correspondence* and *shifts*, is, however, only possible if one lets go of the contrastive outlook – and this was precisely Baker's objective, an objective that has been put into practice in numerous studies comparing translated with non-translated language on the basis of comparable monolingual corpora (e.g. Laviosa 1998, Olohan & Baker 2000, Mutesayire 2004, Xiao 2010, etc.). Although this attention shift towards the target text was a necessary step in the development of TS, voices claiming the inevitability of involving the source text into translational corpus research would quickly be heard too (see section 2.2.2.3). By the turn of the century, CBTS had established itself as a new paradigm within TS:

> This new paradigm, corpus-based translation studies (CTS), can be defined as the branch of the discipline that uses corpora of original and/or translated text for the empirical study of the product and process of translation, the elaboration of theoretical constructs, and the training of translators. CTS makes use of a rigorous and flexible methodology, theoretical principles are firmly based on empirical observations, it uses both inductive and deductive approaches to the investigation of translation and translating, and it encourages dialogue and co-operation between theoretical, empirical, and applied researchers (Laviosa 2003, 45).

In that same 1993 seminal article, Mona Baker proposed a research program for CBTS, which has as its most important task to determine what distinguishes translated text from non-translated text:

> [I]t will be necessary to develop tools that will enable us to identify universal features of translation, that is features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems (Baker 1993, 243).

Although Baker initially proposed six different types of *universals* (1993, 243-245), we will give an overview here of the four *universals* as presented by Baker in her 1996 article "Corpus-based Translation Studies: The Challenges that Lie Ahead". This latter list of four *universals* – each of which now properly named, unlike the list of six *universals* in the 1993 article – has indeed been taken as a standard reference to Baker's *universals* (with only occasional reference to the sometimes more vague terms used in the 1993 article). The establishment of this list is "[b]ased on small-scale studies and casual observation" (Baker 1993, 243), but by virtue of corpus research, Baker hopes to find evidence for the existence or absence of these presumed *universals*. Before such a testing can take place, refined definitions are of course needed. As we will see, this (refined) defining is not an easy task, and even fine-tuned definitions of each of the *universals* can quite easily be refuted.

2.2.2.1 Explicitation

Before Baker posited *explicitation* as one of the presumed features of translated language, Blum-Kulka (1986) had already proposed the Explicitation Hypothesis, claiming that *explicitation* was "a universal strategy inherent in the process of language mediation" (Blum-Kulka 1986, 21). Applied to TS, it then became "inherent in the process of translation", since translation could be considered as one of the ultimate forms of language mediation. Baker, following Blum-Kulka, then defined *explicitation* as follows:

> I take "explicitation" to mean that there is an overall tendency to spell things out rather than leave them implicit in translation (Baker 1996, 180).

Explicitation may consequently be determined by looking at text length (if it is true that things are overall more spelt out in translation, this should lead to an increased text length); or may manifest itself via syntactic or lexical devices. Amongst the numerous studies that were carried out to test the explicitation hypothesis, we can mention Øverås (1998), Olohan and Baker (2000), Olohan (2003), Mutesayire (2004), Puurtinen (2004) and many others (see Kruger 2012 or Zanettin 2013 for overviews of the *translation universals* literature).

Rather than summarizing the plethora of existing studies on the subject, we zoom in on one study on syntactic *explicitation* that was carried out by Baker and Olohan (2000). The study focused on optional *that* in reported speech and concluded that there was indeed an overall preference to use *that* instead of the zero-connective in translated as opposed to original English (the study concentrated on forms of *say* and *tell*) (Olohan &

Baker 2000, 157). Although the evidence and argumentation in favor of this conclusion do seem convincing and are often cited as a confirmation of the *explicitation* hypothesis, Becher (2010, 10-11) argues that the observed increase of optional *that* in translated language can be more plausibly explained as either source language interference or conservatism. As for source language interference, the increased use of *that* may be explained as follows: some source languages may require *that* in reported speech, other source languages may or may not allow it. The source language(s) (if they were known, which is not the case in Olohan and Baker's study) could then explain the increased use of *that* in the sense that the greater the number of source languages in the corpus which require *that*, the more likely the increased number of *that* in translated language is due to source language interference. The increased number of *that* in translated language could also be attributed to translators' alleged conservatism (Becher, Ibid.). If Baker's statements (1993, 244; 1996, 183) that translators have more conservative linguistic habits than other text writers are to be taken as true, Becher argues that it would in fact quite straightforwardly (or at least more straightforwardly than the *explicitation* hypothesis) explain the increased use of optional *that*, since this is the more 'conservative' option in English (it cannot be left out after more formal and fewer common verbs).

Although Baker's definition of *explicitation* seems quite unequivocal at first sight, and (quite) easy to identify contrastively on an individual sentence level, it is much more difficult to maintain it as a universal hypothesis and even less so when the implied source languages are unknown and cannot be taken into account. A phenomenon of zero-attestation vs. attestation may or may not be interpreted as *explicitation*, but, as Becher (2010) has shown, other hypotheses that "do not presuppose a subconscious tendency to explicitate on the part of translators" (p.11) may easily overrule it. Becher, for that matter, also refutes Øverås' (1998) arguments in favor of *explicitation* (Becher 2010, 12-16). He furthermore concludes that translators opt for *explicitation* on the basis of the same considerations as writers of original texts do and that there is consequently no such thing as translation-inherent *explicitation* (Becher 2010, 22-23).

2.2.2.2 Simplification

> We can tentatively define "simplification" as the tendency to simplify the language used in translation (Baker 1996, 181).

The main question is again how *simplification* can be operationalized in a corpus study. Baker suggests that "[t]ranslators [...] may be inclined to break up long sentences in translation, so we might look at average sentence length in both source vs. target texts [...]" (Baker 1996, 181). Laviosa-Braithwaite (1996b) carried out such a study and found that average sentence length in translated texts was significantly lower than average sentence length in a corpus of non-translated texts (Baker 1996, 181). However, the argument that shorter average sentences are ‘simpler’ than longer sentences is a (mere) intuition about how texts can be ‘simplified’. In research related to Second Language Acquisition, it has been shown that coherence markers increase text comprehension more than fragmentation (the use of shorter sentences) does (Land et al. 2009). So, even if it were true that the average sentence length in translated texts is shorter than in non-translated texts, and even if the translators did produce shorter sentences out of a primary concern with the comprehensibility of their text, this does not mean that the text does *de facto* become simpler. Although “simplification involves making things easier for the reader” (Baker 1996, 182), conscious acts to do so may well have a contrary effect. Baker adds that, although *simplification* does not necessarily mean that the text is rendered more explicitly, “it does tend to involve also selecting an interpretation and blocking other interpretations, and in this sense raises the level of explicitness by resolving ambiguity” (Baker 1996, 182). An act of *simplification* may thus be realized via an *explicitation* in the text, which makes it obviously extremely hard for the TS researcher to distinguish *explicitation* from *simplification.*

Another way of operationalizing *simplification* is via indicators such as *lexical variety* or *lexical density*. *Lexical variety* (also called *lexical diversity* or *vocabulary range*) can be accessed via the calculation of the type-token ratio – the number of unique word types per total number of (or usually per thousand) tokens. The closer the type-token ratio is to 1 (or 100%), the more varied the vocabulary in a given text/corpus (see e.g. Laviosa 1998). *Lexical density* (information load) is “the percentage of lexical as opposed to grammatical items in a given text or corpus of texts” (Baker 1995, 237). Different text types can, however, show different levels of *lexical density*, so that the measure can only be used for intra-text type comparison in TS. Alternatively, *lexical density* can be measured by calculating mean word length (Kruger & Van Rooy 2012). The use of this measure is based on the assumption that “word length can be seen as a measure of morphological complexity. [...] mean word length is also an indicator of lexical specificity. Shorter words are more frequent and more general, while longer words are less frequent and more specific” (Kruger 2012, 366).

Contrary to the universal of *explicitation*, it seems quite easy to determine whether *simplification* has taken place on the level of a text or (part of) a corpus. The measures proposed above are quantitative and very little or no doubt can arise as to how to operationalize a type-token ratio or a mean word length. However, one can ask oneself to what extent these measures really indicate *simplification* in Baker’s sense of “making things easier for the reader” (Baker 1996, 182). As mentioned above, some of the measures that are taken into account such as average sentence length do not seem to “make things easier” (Baker, Ibid.) at all. In addition, if we take a look at readability research which is equally concerned with “what makes some texts easier to read than others” (Dubay 2004, cited by De Clercq et al. 2014), we see that the above proposed measures for *simplification* in translated texts do not suffice (any longer). Although traditional readability formulas do or did indeed use the kind of measures proposed above as measures of *simplification*, readability research has evolved rapidly over the last decade or so:

> In recent studies, readability has been linked with more complex lexical and syntactic text characteristics [...] and more recently, discourse features capturing local and global coherence across text are also being scrutinized [...] (De Clercq et al. 2014, 294).

A more up-to-date and complete measure of *simplification* would thus necessarily have to take into account advances made in readability research before any statements could be made as to the overall *simplification* of a text or corpus under study. Although readability measures were used to assess the difficulty of the source text of a translation task (Jensen 2009; Sun & Shreve 2014), measures of readability have – to our knowledge – not yet been used to test this translation *universal* but could give researchers firmer quantitative ground to stand on in the comparison of translated and non-translated texts. A final point which follows out of this concerns the pejorative connotation of the term *simplification*. If one is indeed interested in discovering whether translated texts are indeed easier to understand, we could take the point of view of readability: rather than asking whether translations are 'simpler' than non-translations, we could ask: do we see that factors commonly known to raise readability equally appear in translated texts?

2.2.2.3 Normalization/conservatism

> "Normalisation" (or "conservatism") is a tendency to exaggerate features of the target language and to conform to its typical patterns (Baker 1996, 183).

The third universal feature of translation, also referred to as "conventionalization", "standardization" or "sanitization" (Zanettin 2013, 23) is, according to Baker, "quite possibly influenced by the status of the source text and the source language", in the sense that a higher status of the source language will decrease the tendency to normalize (Baker 1996, 183). Quite strangely, within the first paragraph which defines what *normalization* ought to be like, it already dismisses itself as a universal *strictu sensu* since source language influence is "quite possible". Source-language related phenomena such as *interference* were nevertheless excluded from the *universals* research paradigm, posing "serious problems for any kind of causal explanation of the findings" (Pym 2008, 311). This being said, *normalization* in translation has been widely researched via apparent operators such as hapax legomena as a feature of the lexical creativity (Kenny 2001), typical grammatical features (Hansen-Shirra 2011) and degrees of formality of pairs of near synonyms (De Sutter et al. 2012) etc. The results of these studies, however, are far from straightforwardly stating that *normalization* is 'usual business' in translation. Kenny (2001, 210) concludes that “lexical *normalization* has been found, but it is far from an automatic response to lexical creativity in source texts”. De Sutter et al. (2012, 338) conclude that degrees of formality in translated texts may differ depending on the source text and that translated texts are not always more formal (more conservative) than non-translated texts, thus only partially confirming the conservatism hypothesis.

An important factor which seems to be heavily influencing the results about *normalization* is source language. As Baker had herself intuited about this universal, it seems very hard to pretend that normalizing trends in translated texts are (completely) source language independent. It has indeed been observed that, when translating into the same target language, translators normalize *less* when translating from one source language and *more* when translating from another source language (see De Sutter et al. 2012). Investigating source language influence on translated texts, especially with regards to the *normalization* hypothesis, seemed now almost inevitable. Some researchers, like Teich indeed hypothesized a two-directional influence on translated texts:

> - translations are different from comparable texts in the same language because the *source language shines through*. How does the source language shine through in translations and how can this shining-through be described?
> - translations are different from comparable texts in the same language because they try to be even more ‘typical’, *more ‘normal’* of the target language than are original texts in the same language. In what terms can ‘normal’ be defined and how can that definition be applied to translations? (Teich, 2003: 61-62, our emphasis).

While the second hypothesis straightforwardly corresponds to the universal of *normalization*, the first one follows from the idea that certain phenomena in translated texts can only be explained when the source language is taken into account. In this case, the source language literally *shines through* in translated texts so that certain features in translated texts can only be explained with respect to the source language from which they have been translated. Hansen-Schirra (2011, 136) puts forward the idea that the specific features of translated texts might well be the result of a *hybridization* of *normalization* and *shining through*. The specific features observed in translated texts would then hold a balance between a tendency to conform to the *norms* of the target text and a propensity to adopt features that are typical for the source language at hand.

Tendencies of *normalization* and *shining through* may well exist on the semantic level too. Research questions with respect to semantic *normalization* and *shining through* could then be posited as follows: on a *semasiological* level, we could ask ourselves whether the different meanings of a given source language word (the different concepts it expresses) have an influence on the meanings attributed to / the concepts expressed by a target language word (do they *shine through*?) so that the meanings expressed by the same word (or the prototype-based organization of those meanings) in translated and non-translated texts would differ from each other under influence of the source language; or, will the expressed meanings in translated language conform to or even exaggerate (the organization of) the meanings expressed in non-translated language (*normalization*)? On an *onomasiological* level, we could pose the following question: will the organization of the words expressing a given concept be altered under influence of translation, so that the structure of the source language semantic field *shines through* in the structure of the translated semantic field; or, will, on the other hand, the organization of the words expressing a given concept in translation tend to be more similar (*normalize*), or even exaggerate the structure of non-translated target language?

2.2.2.4 Levelling out

> "[T]he tendency of translated text to gravitate towards the centre of a continuum" (Baker 1996, 184).

While this definition might seem somewhat vague, the idea of *levelling out* means that "we can expect less variation among individual texts in a translation corpus than among those in a corpus of original texts" (Baker 1996, 177). Translated texts would thus be more alike amongst each other than non-translated texts. Just like for *simplification*, the measures to investigate *levelling out* are lexical density and type-token ratio; the difference lies in the conclusions that are drawn from these measures. From the point of view of *simplification*, a lower lexical density in translation leads to the conclusion that translated texts are more simple than original texts. Seen from the perspective of *levelling out*, the question is raised whether the lexical density amongst translated texts is more similar than lexical density amongst non-translated texts. In other words, *levelling out* is investigated by comparing the variation of a certain feature (e.g. lexical density or type-token ratio) between translated and non-translated texts (Baker 1996, 184). As Baker already indicated in 1996, *levelling out* is probably the universal that has received the least attention in the literature. Olohan's 2004 overview of the state of the art in corpus studies in translation confirms that this universal is the one for which least empirical investigation has been set out as it seems to be the most difficult one to measure (Olohan 2004, 100). Later overviews by Kruger (2012) or Zanettin (2013) show that the decade following Olohan's overview has not brought much change to this. Kruger mentions the existence of the universal of *levelling out* but does not take it up in her overview of *universals* (most probably because there were no studies focusing on *levelling out* to mention). She does indicate, with respect to her own study presented in the same article, that some evidence has been found for this universal "since register differences are largely neutralized in the translated subcorpus" (Kruger 2012, 369).

Zanettin mentions not one study investigating “linguistic indicators of leveling out or the way to implement them through computational operators” (Zanettin 2013, 23).

In short, although the idea behind the universal of *levelling out* is potentially interesting, no studies have so far focused on this universal in particular, most probably because it is often (mis)taken for the universal of *simplification,* a universal that is operationalizable ‘on the surface’ of the corpus and does not require the use of statistical techniques- which are needed if one wants to gain more insights into the *levelling out* of a certain feature. In order to arrive at an understanding of *levelling out*, one would indeed need to have an idea of an average range of a specific feature in translated texts and compare it to the average range of that same feature in original texts so as to understand whether translated texts are more like each other than non-translated texts. It is, in our opinion, precisely investigations into meaning in translation that would best ‘suit’ this universal (which would consequently explain why this *universal* has never been properly investigated). Contrary to the other *universals*, a certain level of abstraction will be needed (a common requirement for both the investigation of meaning as well as for the universal of *levelling out*).

The following research questions may then apply for semantic *levelling out*. On a *semasiological* level, would one expect a given word in translated language to show less meaning differentiation (fewer different meanings implied or an alternation in the structure of those meanings) than for the same word in non-translated language? On an *onomasiological* level, would a given concept in translated language be expressed by fewer lexemes (a restricted semantic field), or would the organization of the lexemes in the field be influenced by the translated linguistic status? Arguably, this type of question could also be claimed to fall under the universal of *simplification*, but our decision to consider it here as *levelling out* has three reasons: (i) it is our purpose to compare the semantic variation of translated language to that of non-translated language, and in that respect, our purpose corresponds more to the idea of *levelling out* (ii) we want to avoid an association with the negative echo that the universal of *simplification* has in our view, (iii) we want to show that there are ways in which *levelling out* can be measured, something that has not often been done, and this on a semantic level.

2.2.2.5 Universals: the more the merrier?

Although it sounds a little irreverent, we could say that a lot of ink has been spilt over the *universals* in the last two decades. Doubt about the existence of *universals* led Mauranen (2004, 1) to point out that claims of evidence for the existence of *universals* by some researchers, and statements by others that *universals* could simply not be investigated, had left the research community with the existential question whether *universals* did or did not exist at all.

Some *universals* were indeed refuted (see e.g. Becher 2010), others such as the Unique Items Hypothesis (Tirkkonen-Condit 2004) or the Asymmetry Hypothesis (Klaudy 2009) were added to Baker's list. The Unique Items Hypothesis states that some features that are unique to the target language will appear less or not at all in translated language, because they are not triggered by any source language feature (Tirkkonen-Condit 2004, see also Chesterman 2004 for a revision of the hypothesis). The Asymmetry Hypothesis, first proposed by Klaudy (2009) and modified by Becher (2010) affirms that "[o]bligatory, optional and pragmatic explicitations tend to be more frequent than the corresponding implicitations regardless of the SL/TL constellation at hand". Although a *universal* ought to be a feature that appears irrespective of the source language, scholars quickly understood that – in order to figure out where certain phenomena were coming from – the inclusion of the source language seemed inevitable. However, the inclusion of the *a priori* excluded feature *source language* within the *universals* paradigm as well as the expansive number of *universals* was not without consequences for the viability of the notion.

Chesterman proposed to divide the (growing number of) *universals* into two categories, the S-*universals* ("characteristics of the way in which translators process the source text" (Chesterman 2004, 39) and T-*universals* ("characteristics of the way in which translators use the target language"). Amongst the potential S-*universals* (Ibid., 40), we find features such as lengthening (translated texts tend to be longer than their originals) - proposed by Vinay & Darbelnet (1958, 185) – and Toury's (1995) laws of interference (source text features are transferred to the target text) and growing standardization (a source-text specific feature will be replaced by a more 'common' expression in the target text), the latter having a lot in common with Baker's definition of *normalization*. Amidst the potential T-*universals*, we find, *inter alia, simplification* and the Tirkonnen-Condit's Unique Items Hypothesis. Chesterman moreover counters the difficulty of testing *the universals* of translation within the scope of *one* study, by proposing what he calls "the low road", where "a universal hypothesis might also be tentatively proposed on the basis of empirical results pertaining only to a subset. [...] [T]he criteria on which the subset is defined [...] [will] define the conditions that determine and limit the scope of the claim" (Chesterman 2004, 40). We share this idea with Chesterman, in that any of the generalizations made on the basis of a corpus study first and foremost apply to the language-pair(s), the period, the text genre(s) etc. which are selected by the researcher (by selection of the (sub-)corpus that will be used or by the parameters that were set for corpus creation). In order to make general, *universal* claims, the same study would have to be repeated for an as wide as possible variety of language pairs, as many periods and as many genres as possible. The apparent unfeasibility of doing the latter has led researchers such as Mauranen to dismiss the idea of universality at once, but to rather opt for "general tendencies" (2008, 35):

> The term 'universals' does not, then, necessarily refer only to absolute laws, which are true without exception. Rather, most of the suggested universal features are general or law-like tendencies, or high probabilities of occurrence (Mauranen 2008, 35).

It becomes clear that it is precisely the claim of *universality* of the *translation universals* which has been bothering the research community interested in the subject. The newly added *universals* or revisions such as Chesterman's S- and T-*universals* all indeed seem to try to do away with this idea of universal applicability, others rather opt for terms such as *general tendencies* of translation in an attempt to tone down or at least nuance the universality claim. Mauranen refers to the field of general linguistics, where the term *universals* is also used, but where it has become general practice "to take into account different kinds of general tendencies shared by a large number of languages, not only 'absolute' universals, that is, features shared by every human language" (Mauranen 2008, 35), and she suggests that the term *universals* should be defined in a similar way within TS.

As a result of this unceasing *universals* debate, it was realized that translational behavior is *multidimensional* in nature (De Sutter 2013), and that, in addition to purely linguistic matters, there are also a number of social, cultural, ideological and cognitive constraints acting upon translation (Baker 1999). In order to include these constraints into the research paradigm, alternative methodological approaches have recently been proposed. In translation process research, *triangulation* (the combination of several data gathering techniques and methodologies) has become increasingly common (see e.g., Alves 2003; Carl 2010; Hansen 2010). In addition, the use of multivariate statistics has recently been introduced into CBTS. "[M]ultivariate data are typically represented in a matrix form with rows holding the units and columns holding the variables" (Jenset & McGillivray 2012, 302). By representing corpus data as (frequency) matrices, the complexity of the (type of) linguistic data in (corpus-based) translational research can be (more easily) tackled. These techniques indeed "allow us to preserve the rich diversity of linguistic forms, while at the same time reducing the variation in a principled way to a simpler, more interpretable structure" (Jenset & McGillivray, 2012, 301). Recent studies by Delaere et al. (2012); Diwersy et al. (2014) and the studies presented in the edited volumes by Oakes & Ji (2012) and De Sutter et al. (forthcoming) have shown that multivariate statistical methods can be successfully implemented into CBTS. Our own study will equally make use of multivariate statistics to capture the complexity of the meaning relationships in translated and non-translated language by using translations as the variables of source-language lexemes (and vice-versa) in frequency matrices (this will be further explained in section 2.4.2 as well as in chapter 3).

2.2.2.6 The relationship between universals and meaning

The impact of Baker's 1993 article on the development of CBTS can hardly be overestimated. In section 2.2.2 we have claimed that Baker's research program was both necessary and useful for the emancipation of CBTS and even for TS as a whole. However, some of Baker's propositions have heavily determined the focal points of TS in the years to follow. For instance, Baker's dismissal of the source language, in an attempt to put translation and translated language at the center of attention, has led to studies which completely leave out any consideration regarding the source language (since the type of corpora that were favored – comparable corpora – did not include the source texts of the translations in the corpus). This probably also led to an increased amount of comparable corpora (instead of parallel corpora) because precisely these type of corpora were thought to serve the needs of CBTS best, a phenomenon that in its turn led to more target-oriented research in TS.

A similar scenario might apply for the study of meaning in CBTS. By announcing "the decline of the semantic view of translation" (Baker 1993, 237), Baker attempted to get rid of clichéd, simplistic ideas about translation (the idea that translation is a mere word-for-word or sentence-for-sentence contrastive operation), but in this way she also declined to some extent the further study of meaning in translation. In the same way as the concepts of *equivalence*, *correspondence* and *shifts* were dismissed because they were thought to betray a preoccupation with practical issues in translation (Baker 1993, 235), the (contrastive) study of meaning in translation was equally set aside because such a study would imply that the researcher was "still trying to justify them [translated texts] or dismiss them by reference to their originals" (Baker, 1993, 235). Baker's assertions seem to have impacted CBTS in the sense that studies of meaning proper in CBTS are rather scarce, and concepts such as *equivalence*, *correspondence* and *shifts* were absent (because considered unnecessary) from the investigations that claimed to fall under the scope of the research program. This does not mean that there are no studies at all within the scope of CBTS that address the problem of meaning. We will see (in section 2.2.3), however, that such studies are inevitably confronted with concepts such as *equivalence*, *correspondence* and *shifts* once again or that they avoid to engage in research into *universals* of translation.

A second reason why research into meaning in translation might have been shoved aside is that meaning finds itself at the very core of what translation *is:* according to numerous scholars in the field, meaning is "the invariant of translation" (Klaudy 2010, 82). Indeed, "it seems to be firmly embedded in public opinion that in translation it is the meaning that has to remain unchanged" (Klaudy 2010, Ibid.). It appears to be widely accepted that 'invariant meaning' is conveyed via lexicalized expressions from one language to another through translation. However, if we question the invariability of the invariant, we somehow remove the firm ground on which a lot of research in translation has so far been built. Nevertheless, if we want to know if it is true that meaning *is* the invariant of translation, we will necessarily have to engage into empirical research into meaning and meaning relationships in translation.

Finally, a third reason why meaning might not have received the attention it 'deserved' in TS, is that meaning is an abstract notion and therefore difficult to capture. This might have discouraged TS scholars and refrained them from taking up the subject. In the following section 2.2.3, we will have a closer look at some important investigations of meaning in translation that have been carried out over the last two decades in an attempt to understand how TS scholars have dealt with the study of meaning.

2.2.3 The cognitive turn in translation studies

2.2.3.1 Translation and Meaning series

As we have mentioned before, meaning is at the core of what translation *is*, and this becomes immediately apparent if one looks at the multitude of subjects in the ten parts of the Translation and Meaning series (each of the volumes contains the proceedings of an international duo-colloquium which is held every five years in Maastricht and Łódź, under the auspices of Marcel Thelen and Barbara Lewandowska-Tomaszczyk). There does not seem to exist a single branch of TS which is excluded from the series: anything from corpus work over machine translation to dictionary compilation, the translation of literary and holy texts, terminology, translator and interpreter training…; it *all* has to do with meaning. This does not mean, however, that the *invariance of meaning* in translation is questioned, nor that a descriptive or explanatory viewpoint is adopted, and even less so that meaning is considered 'linguistically'. Many studies in the Meaning and Translation series indeed do not take a linguistic viewpoint on meaning and fit the subject of the series because of the general acceptance that translation *is* meaning, even more so the *invariant* of translation which leads to the possibility of including virtually any branch of TS into the series. An article such as Laviosa-Braithwaite's "Comparable Corpora: Towards a Corpus Linguistic Methodology for the Empirical Study of Translation" in the third part of the series (1996b) for instance – although uncontestably making an important contribution to the propagation of corpus research and the *universals* program in TS – does not at all engage with the question of meaning invariance in translation, but implicitly accepts it as a bottom line. Other studies, such as Snell-Hornby's (1992) "Word against Text. Lexical Semantics and Translation Theory" (in the second part of the series) express their interest in lexical semantic studies and the possible contributions linguistics can make to TS, although they do not take the viewpoint of the TS scholar investigating translation but instead consider lexical semantics in view of its usefulness for the professional translator (Snell-Hornby 1992, 100). Although a number of studies in the Translation and Meaning series use corpora to investigate (linguistic) meaning, they often focus on the *utility* of such a study in light of translation teaching and translation quality assessment (e.g. Bednarczyk 1997; Lan & Bilbow 2007; Oster & van Lawick 2008) and do not challenge the idea of meaning itself. One of the few studies that actually does engage into the question of meaning *invariance* in translation is Halverson's "Norwegian-English Translation and the Role of Certain Connectives" (1996) where connectives are classified according to semantic categories which are subsequently compared. Halverson concludes that connectives change their semantics in translation and in this way, her conclusions point in the direction of the possibility of *variation* of meaning in translation.

2.2.3.2 The re-introduction of meaning in translation studies

The corpus turn from the 1990s – which ensued from the elaboration of a descriptive approach to translation studies called descriptive translation studies (hence: DTS) with the work of Toury (1980), Hermans (1985, 1999) and colleagues – was the necessary prerequisite for translation studies to become a discipline in its own right. Next to corpus research, Lewandowska-Tomaszczyk (2002) discerned "cognitive approaches to language" as one of the two "leading recurrent themes" (p.41) in translation theory in the 1990s. Other scholars such as Boase-Beier equally stated in the early 2000s that a cognitive turn was taking place in TS (Boase-Beier 2006)[4]. Since in cognitive linguistics, the main emphasis lies on the status of meaning (Lewandowska-Tomsaszczyk 2002, 41), one would expect the so-called cognitive turn to revive an interest in (linguistic) meaning within TS. However, the earliest attestations of a cognitive translation studies were not immediately showing interest for a re-introduction of linguistic meaning (and *equivalence*) into TS but were rather focused on corpus-based empirical and experimental (process) research such as think-aloud protocols for example (see e.g. the volume edited by Tirkkonen-Condit & Jääskeläinen 2000) which equally benefited from the cognitive-translational setting. Rojo and Ibarretxe-Antuñano (2013) summarized what was happening on the intersection of cognitive linguistics and translation by the end of the 1990s as follows:

> The relevance of cognitive linguistics for translation arises mainly from the "experiential" notion of meaning proposed by cognitivists, which abandons the traditional notion of referential truth and highlights the central role of human experience and understanding (Rojo & Ibarretxe-Antuñano 2013, 7).

[4] In the case of Boase-Beier, cognitive insights into TS were thought to have the ability to bridge the gap between literary and non-literary translation.

Although the "notion of referential truth" – referring to *equivalence*-like conceptualizations – was abandoned, a number of scholars did, however, focus on the "linguistic-cognitive orientation" (House, 2013). Early attestations of this linguistic-cognitive orientation include Tabakowska (1993), who proposed to introduce notions from Langacker's cognitive Grammar in TS, and Kussmaul's (1995) idea that *foregrounding* and *suppression* of semantic features could be useful when translating complex meanings (Rojo & Ibarretxe-Antuñano 2013, 8). Research by Wilss (1996) also pointed towards a cognitive-linguistic approach to meaning in the 1990s. House makes a strong plea in favor of a "linguistic-cognitive orientation" in TS, since in her opinion "translation is above all an activity involving language and its cognitive basis" (House 2013, 47). She further argues that TS has been so pre-occupied with "external social, cultural, personal, historical etc. factors impinging on translation 'from the outside' " that it seems to have been missing "the point about the essence of translation" (Ibid.). A cognitive view on translation is then both insightful and necessary in order to "describe and explain how strategies of comprehending, decision making and re-verbalization come about in a translator's bilingual mind" (Ibid., 46).

Since our own study takes precisely this linguistic-cognitive view on translation, we will focus in the following paragraphs on research which has specifically engaged with the linguistic-cognitive orientation on TS. Our overview does not intend to be exhaustive, but focuses on those studies which in our opinion have provided innovative insights for the study of translation within a linguistic-cognitive framework.

Klaudy (2010)

Klaudy (2010) proposes to take the point of view of *translation universals* to investigate phenomena such as lexical specification and generalization. Her contrastive view on translation (translation as transfer) is in apparent opposition with the mainstream research into *universals* (which refutes the contrastive concepts of *equivalence* and *correspondence*). Klaudy shows that the introduction of an *equivalence*-like concept (although a narrow one in comparison to what most translation studies scholars would understand as *equivalence*) makes it possible to study *universals* in a contrastive setting (including both source and target language into the equation). She introduces the concept of lexical transfer, which covers "all the systemic and routine-like operative moves developed by generations of translators to handle the difficulties stemming from the different lexical system and cultural context of the two languages functioning together in the process of translation" (Klaudy 2010, 81). She argues that such a concept is useful in order to explore language differences distinct from those foreseen by contrastive lexicographers and bilingual dictionary builders and she expresses a special interest in "how these systemic differences are brought in motion in the process of translation" (Ibid., 82). She is particularly interested in what happens to meaning in translation and questions the “firmly embedded idea” that meaning remains unchanged in translation (Ibid.). Klaudy’s explanation of what happens in translation is based on a distinction between meaning and sense, where meaning is “the criteria for the usage of a linguistic sign within a given language” and sense “the relationship between the linguistic sign and a certain segment of reality (objects, events, persons, phenomena) here and now, i.e. an actual relationship becoming manifest in a certain communicative situation” (Klaudy 2010, 83). The latter relationship is recreated in the TL “instead of retaining the SL meaning” (Ibid.). Klaudy sees the fact that translators “try to relate TL signs to reality according to SL rules of usage” as a frequent source of translation errors (Klaudy 2010, 83). Although we are more interested in the differences between semantic relationships in translated language and non-translated language – rather than in a contrastive comparison of systemic differences between languages – we will nevertheless follow Klaudy’s proposition to take the initial point of view of *translation universals* to investigate semantic differences.

Martín de León (2013)

In the volume edited by Rojo and Ibarretxe-Antuñano (2013), Martín de León explores how cognitive models of meaning can be of use for translation. She states that “different cognitive approaches provide different visions of meaning, and that they also lead to different theoretical frameworks for empirical translation research” (Martín de León 2013, 99). If meaning is seen as invariant (“transferable, invariable information units”), then, the task of the translator can be resumed to a transferring of information encoded in the source language into the target language (Ibid.). If, however, meaning construction is seen as *variant* (“a complex, dynamic, and situated process”), the translator’s task will consist in seeking “to provide target readers with the tools they need to construct their own meaning in their own situation (Risku 2004)” (Martín de León 2013, 99). Since in translation, “meaning construction processes are partly artificially situated (Holz-Mänttäri 1990), [...] it provides a particularly interesting arena to empirically research these processes, where [...] different cognitive approaches to meaning construction can be tested” (Martín de León 2013, 99.). In this article, Martín de León expounds different cognitive paradigms and shows how some paradigms such as the “classical paradigm” which relies on the idea that “symbols of mental language are abstract, amodal and arbitrary” (Ibid., 100) - are unable to explain the processes involved in human translation (Ibid., 103)[5] while others such as distributed cognition (Ibid., 115) are better fitted to explain the processes of human translation. The two

[5] This rejection of the classical cognitive paradigm was already introduced by Risku (1998), who, in later publications (Risku 2002, 2004) proposed the distributed cognition approach.

starting points of this discussion – (i) translators provide target readers with tools to construct their own situated meaning, and (ii) meaning construction processes are partly artificially situated in translation – are, however, debatable. If a translator chooses to translate the English lexeme *house* into Dutch *huis*, he does not "provide the reader with a tool to construct his own meaning" but merely 'imposes' the Dutch meaning of *huis*. Like any mention of *huis*, be it in a translated or a non-translated text, the reader's understanding of *huis* will be mediated by his own embodied experience. It seems difficult to believe that such meaning construction by the reader would be more 'artificial' because of the translational status of a text. However, the construction of the meaning of *huis* by the translator – in the translator's bilingual mind – can possibly carry features of both *house* and *huis* – and is in this sense partly artificial. If this latter understanding of the construction of meaning is indeed intended, then, a cognitive-empirical account of meaning construction in translation which uses a representational model will be needed. Rather than accepting the artificiality of meaning construction in translation and taking it as a starting point to test various cognitive models – which is what Martín de León does – our study will set out to investigate how meaning relationships may differ between translated and non-translated language. Yet, this requires us to choose one (or several) cognitive view(s) on meaning as a starting point (since this will determine our way of visualizing the meaning construction in translation). In this study, we will adopt a prototype-based view on meaning, engendering representations of meaning that are built up around a prototype (see section 2.4.3).

Korning Zethsen (2008)

One study which explicitly links corpus-based cognitive (lexical) semantics to the study of translation is conducted by Korning Zethsen (2008). She proposes to use cognitive semantics "as a tool for researchers within translation studies (TS) who are particularly interested in revealing evaluative aspects of the units of meaning of source texts and their translations" (Korning Zethsen 2008, 249). In her view, meaning arises through interpretation. Information provided by the linguistic expressions is combined with contextual information and in this way triggers different interpretations (Korning Zethsen 2008, 250). Since "certain words presuppose a certain context to such an extent that this context can be said to form part of the lexical meaning of the word [...] individual word meaning cannot be considered a sound concept within semantic analysis" (Ibid.). This assumption entails that the *unit of meaning* has to be extended – to a phraseological *unit of meaning* (Sinclair 1996). If this is true for the *unit of meaning*, then it is equally true for the *unit of translation* (Korning Zethsen 2008, 250).

Before we further explore Korning Zethsen's investigations regarding the contribution of lexical semantics to TS, a small note on *units of meaning* and *units of* *translation* is in order. Just as linguists are in constant search to delimit *units of meaning*, TS scholars have transposed this question to the study of translation and the *unit of translation*. First introduced by Vinay and Darbelnet (1958) as *unité de pensée*[6], the debate of what exactly constituted the "smallest segment" that could not be translated separately has, just as the debate on *units of meaning*, been ongoing and is often defined in the light of the scholar's view of what translation is and how it should be investigated. First, whether the scholar who defines the *unit of translation* adopts a process-oriented or rather a product-oriented view on translation will impact his definition of the *unit of meaning*. Second, the scholar's view on translation – does he consider to investigate translation on the cultural, the textual, the phraseological or the word- or morpheme level – will equally influence the way in which he defines the *unit of translation*. The debate of what exactly constitutes the *unit of translation* is, however, "often conducted in terms of a strict opposition between translating word-for-word and translating sense-for-sense" (Laviosa-Braithwaite 2004, 286). The way in which the *unit of translation* is defined by the researcher thus becomes an ideological choice tending to prescribe how one (the scholar, the translator, the reader) should consider the unit/should take up the translation task. A concept that is distinct from but closely related to the *unit of translation* is *equivalence*. Laviosa-Braithwaite (2004, 287) confirms that "[i]t is clearly possible to establish EQUIVALENCE between units smaller than the clause even when it is clear that the clause is the unit of translation". *Equivalence* can then exist on a sub-translation-unit level, which, in our view, makes *equivalence* – in comparison to *unit of meaning* – a better discernible, less ideologically determined candidate to study translated language in comparison to non-translated language.

Korning Zethsen (2008) takes a prototype-based view on meaning "which in addition to inherent lexical meaning helps us account for and describe evaluative meaning which is not necessarily inherent in the lexeme" (Korning Zethsen 2008, 251). She acknowledges that practical reasons may force the researcher to work on the level of semes and suggests that scholars "should aim at a description of prototypical features, inherent or contextual" rather than "attempting an exhaustive analysis of a lexeme" (Ibid.). She proposes to focus on semantic prosodies, i.e. "the spreading of connotational colouring beyond single word boundaries" (Partington 1998, 68 in: Korning Zethsen 2008, 256) which have not often been investigated by contrastive comparison. She concludes that:

> Semantic prosody is bound with time to influence our perception of the concept of *equivalence*. A likely hypothesis is that the traditional problem of 'false friends'

[6] "le plus petit segment de l'énoncé dont la cohésion des signe est telle qu'ils ne doivent pas être traduits séparément" (Vinay & Darbelnet 1958, cited by Nord 1997, 68).

> within translation is much more pervasive than assumed up till now. Presumably equivalent words may have developed differently in two languages and have in time been influenced by the company they have kept and thereby developed different prosodies (Korning Zethsen 2008, 258).

Korning Zethsen touches here upon a matter that might well be pervasive in translation, but which needs advanced (corpus) methods to be revealed. Moreover, by putting a concept such as semantic prosody at the center of attention of translational research, not only does she re-introduce the concept of *equivalence*, she also questions the 'invariance' of meaning in translation and shows how corpus-research accompanied by interpretation can be used to uncover the importance – Sinclair (1997) has argued that semantic prosody might well be the first determinant of word choice – of such subtle issues as semantic prosody in translation.

Halverson (2003, 2010, 2013, forthcoming)

One of the few scholars who has been consistently occupied with the study of meaning in translation is Halverson (2003, 2010, 2013, forthcoming). Since the beginning of the 2000s, she has been developing a hypothesis that could account for the observed differences – allegedly due to some kind of translational effect – between translated and non-translated language, from a cognitive perspective. She asserts that *translation universals* possibly have a cognitive basis, i.e. that they "arise from the existence of asymmetries in the cognitive organization of semantic information" (Halverson 2003, 197). Halverson is convinced that cognitive linguistic theories can inform TS in such a way that they can possibly provide explanations for the generalizations that are empirically accounted for in TS, i.e. (some of) the *universals* (Halverson 2003, 230). She proposes a hypothesis, the Gravitational Pull Hypothesis, which combines Langacker's (2008) Cognitive Grammar with De Groot's (1992) theory of bilingual semantic representation (see chapter 5 for a more detailed account). Shortly put, patterns of over- and underrepresentation which are observed in translated language are thought to be due to particular patterns in bilingual semantic networks, with higher or lower activation of certain patterns leading to more or less selection of that particular pattern. Some patterns exert some kind of a pull; pushing (or rather, pulling) the translator to use a certain target (lexeme, expression, structure) more (or less) prominently than another one. In chapter 5, we will explain how specific characteristics of the bilingual schematic network can lead to over- or underrepresentation of certain features in the network. For the time being, suffice it to know that the Gravitational Pull Hypothesis was conceived to give explanatory value to the generalizations uncovered by the *translation universals*. The hypothesis creates possibilities to investigate questions of meaning within TS and proposes to do so via the mapping of schematic networks. This view corresponds indeed to our idea to visualize meaning in translated as opposed to non-translated language so as to uncover differences and similarities on the semantic level between the these varieties. The GPH will however not be of help for the methodological development that allows for such visualizations, but will be of primary importance when it comes to explaining the observed phenomena in the schematic networks (see chapter 5).

Conclusion

The studies that were presented in this section all deal with the concept of meaning within a cognitive-linguistic view on translation. They all agree on the important point that the re-introduction of meaning research into translation is not considered as an obstacle to the study of *universals* (some even claim that they could well provide explanatory hypotheses for these *universals*).

Obviously, the contributions of the authors cited above reach further than this. The research of each of these scholars has in fact contributed in various ways to the study of meaning in translation. Firstly, Klaudy emphasized the need to re-introduce an *equivalence*-like concept for the contrastive comparison of systemic differences between languages in translation, which in fact allows for a re-inclusion of the source texts into the comparisons. Although our ultimate research goal is not a contrastive comparison (but rather an *intralingual* comparison between different varieties of Dutch, i.e. translated and non-translated Dutch), the contrastive method(s) that we plan to rely on equally necessitate(s) the re-introduction of such an *equivalence* concept. Secondly, Martín de León drew our attention to the importance of acknowledging the possibility that meaning might in fact *not* be invariant translation. She further showed that different theoretical frameworks may be applied to empirical translation research. Thirdly, Korning Zethsen explicitly linked corpus-based cognitive (lexical) semantics to the study of translation. In that regard, her methodological intentions are closely linked to ours. Convinced about its descriptive properties, she proposed to take a prototype-based view on meaning, which is also the viewpoint taken in this study. Her research into semantic prosody further puts into question the notion of the invariance of *equivalence*. Finally, Halverson clearly explains the possibility that *universals* have a cognitive basis. Her Gravitational Pull Hypothesis implies that specific characteristics of schematic bilingual networks may have translational effects. Halverson suggests that the study of meaning structures might in fact open up ways to explain a number of phenomena that have (since long) been observed in translation. In chapter 5, we will use the Gravitational Pull Hypothesis to explain some of the phenomena that emerged from the comparison of semantic fields of translated and non-translated language.

2.2.4 On a tightrope with equivalence

The notion of *equivalence* is one of the most heavily loaded concepts in translation studies. A number of developments within the discipline – ranging from Nida's socio-linguistic translation analysis (Nida 1964; Nida & Taber 1969) to skopos theory (Nord 1997) and including cultural, power and other turns – went to show a gradual but consistent attention shift from the individual word *equivalence* level to a more holistic view on translation (Munday 2009, 10). However, throughout the last forty years or so, no real consensus was reached on the concept of *equivalence*. Early linguistic approaches – think of Vinay and Darbelnet's *Stylistique comparée du français et de l'anglais* (Vinay & Darbelnet, 1958) for example – were often disregarded as they were said to narrow down the scope of translation to mere transcoding (Vandeweghe et al. 2007, 1) whereas historical-descriptive studies of translation as well as many of the early studies within the *universals* paradigm – which generally concentrated on the target text – made the need for a contrastive concept such as *equivalence* disappear *de facto*[7].

A linguistic-oriented study of translation such as ours which takes into account both source and target language will nevertheless need a solid definition of the concept of *equivalence*; it is impossible to dismiss the concept in a study which will rely on and investigate contrastive relations between source and target language. Because of this linguistic-cognitive view on translation, and in view of formulating our own definition, we are particularly interested in how the 'early' linguistics-oriented scholars defined *equivalence*.

In his work 'A Linguistic Theory of Translation' (1965), Ian Catford differentiates between *equivalence* as a (contrastive) empirical phenomenon "discovered by comparing SL and TL texts" (Catford 1965, 27) and the idea that one can or should 'justify' *equivalence* by discovering its underlying conditions. This distinction is an important one because it shows that although the underlying conditions that justify *equivalence* may be complex and cause of debate, the notion itself need not be problematic, provided that one 'solely' considers *equivalence* as an empirical phenomenon. In the 1970s, the word-phrase *equivalence* level was gradually abandoned and *equivalence* was sought on the textual level (see e.g. Koller 1979). The source language orientedness of *equivalence* was, however, not questioned. The problem with the early linguistically-

[7] This does not mean that all scholars have dismissed the equivalence concept; see e.g. Pym (2007) who identifies the difference between "natural" and "directional" equivalence as one of the causes of misunderstanding about the equivalence concept and re-introduces this distinction to interrogate contemporary localization projects (Pym 2007, 271).

oriented idea of *equivalence* seemed thus to reside in the source-oriented as well as the (innate) prescriptive nature of the *equivalence* concept.

In the early 1990s, Reiss' and Vermeer's *Skopos* theory (1991) lays the emphasis on the *purpose* of a translation and *equivalence* becomes "one possible relationship among others" (Schäffner 1999, 5). Toury takes this idea one step further, and states that *equivalence* is "any relation which is found to have characterized translation under a specified set of circumstances" (Toury, 1995, 61). Toury's notion of *equivalence* (1980, 37 ff.) is to a large extent based on Catford's definition to which he adds the notion of relevance: "relevance for ST [source text], or from ST's point of view, does not imply relevance for TT [target text], or from TT's point of view". Translation *equivalence* is thus defined differently depending on the point of view one takes. From the source text's point of view, *translation equivalence* equals "the "similar relevant features" which both source text and target text are "relatable to" (Toury 1980, 38), whereas from a target text's point of view, *translation equivalence* is "an empirical fact [...] the actual relationships obtaining between TT and ST" (Toury 1980, 39). Toury further notes that in this type of description the term *equivalence* is used in two different senses: as a theoretical term (which then refers to an "abstract, ideal relationship" and as a descriptive term (referring to "actual relationships between actual utterances in two different languages"). The fact that within one description, *equivalence* can carry both senses shows, according to Toury "a discrepancy, even a gap, between theory and actual phenomena, or between theory and the possibility of accounting for this phenomena" (Toury 1980, 39). He further adds that it is "precisely this gap which so clearly indicates the inadequacy of a source-oriented theory of translation to serve as a basis for the study of phenomena, actually belonging to the target pole" (Toury 1980, 39).

In sum, both Catford and Toury claim that one of the possible ways of defining *equivalence* is to consider it as the observed/empirical relation between source and target language. Toury explicitates that a specification of what this relationship 'should' be stems from a theoretical, abstract idea of *equivalence* which is incompatible with the idea of *equivalence* as an empirical relation. If we accept *equivalence* as the observed/empirical relation between a source and a target language entity, and abandon the theoretical, source-oriented definition of *equivalence*, it consequently becomes possible to investigate this relation and to comprehend *post-hoc* what this *equivalence* is made of (rather than impose an *a priori* theoretical and idealized *equivalence* notion).

Within a corpus study, the observed relation between a source and a target language entity is implied by the corpus alignment, i.e. whenever man or machine establishes an alignment between two linguistic entities, this alignment implies that the two contrastive linguistic entities are considered equivalents without this statement implying any value judgment on the content of the *equivalence* relation. Such type of *equivalence* is established *post-hoc* – contrary to a prescriptive *a priori* definition of *equivalence*.

We now propose to define equivalence as follows: the *equivalence* relation exists when one expression in the target text is recognized as a translation of a source language expression or when one expression in the source text is recognized as the source language expression of a translation. This identification does not further engage into any value judgment about the relation itself between the source language expression and the translation. Our definition does not impose any prescriptive 'rule' on what is acceptable or not as *equivalence*, is bi-directional (meaning that it can be established by looking first at the source text and then at the target text, or vice-versa) and can hold on several levels (word/phrase/text). This definition is indeed greatly indebted to Catford and Toury's idea of *equivalence* as an empirical relation. Rather than imposing on the *equivalence* relation a need to be "the closest natural equivalent", in our view, *equivalence* can be thought to represent the *relation* between the source and the target text, that what *binds* source and target, irrespective of the nature of *what* is represented in this binding relation. This definition forms the baseline of our idea of *equivalence*. This suffices for now, but we will see that the operationalization of the *equivalence* concept for the purpose of this study will require an extremely pragmatic definition of *equivalence* so that it can be applied to a manual word-level annotation procedure of a sentence-aligned corpus (see section 3.3).

2.2.5 Conclusion

In this section, we have seen that the study of meaning relations in translation is still largely unexplored. Within the most well-known paradigm of CBTS, the *universals* of *normalization-shining through* and *levelling out* appear to be good candidates for the study of semantic differences in translated and non-translated language. We have equally tried to provide a definition of *equivalence* that can be operationalized in an empirical corpus study such the present one, a necessary step if we want to investigate meaning in translation. Although we have formulated a practicable definition of *equivalence* in this section, we indeed still need to take the step towards operationalizability of the notion of (translational) *equivalence*. However, very few studies have suggested and even less so actually developed methodological procedures to do so for research into meaning relationships in translation. In the next section, we will therefore explore some contrastive corpus studies who have engaged with the notion of *translation equivalence* and have proposed valid ways of operationalizing it.

2.3 Contrastive corpus studies

In this section, we will focus on corpus approaches that have manifested an explicit interest in the contrastive study of meaning via corpora. Our goal is to find a way in which a tool of contrastive analysis (an *equivalence*-like concept) can be used in such a way that it is acceptable for a translational analysis, without 'violating' the nature of its subject of research. The question we want to keep in mind throughout this section is how a linguistically inspired notion of *translation equivalence* can be used in such a way that it meets the following requirements. Firstly, the adopted notion of *translation equivalence* needs to allow us to compare translated to non-translated language. Since we adopt a TS point of view in this study, we consider translated and non-translated language as different varieties; we will therefore need to find a way to distinguish between translated and non-translated language. Secondly, whenever a relation of *translation equivalence* is established, we need be sure that it conveys meaning but that the relation itself will furthermore not imply that the conveyed meaning is invariant. We will therefore explore a number of studies which have operationalized *translation equivalence* in contrastive research settings. In this section, we will first focus more generally on the use of translations in contrastive studies (section 2.3.1), before we focus on the procedure of *back-translation*, which is considered as one (of the most) fruitful applications of translation in a contrastive context (section 2.3.2). We will pursue by exploring two successful applications of *back-translation* in contrastive analysis: Mutual Correspondence (section 2.3.3) and Semantic Mirroring (section 2.3.4).

2.3.1 Use of translations in contrastive studies

The close relationship between translation studies and contrastive linguistics and the different types of cross-fertilization(s) that exist between the two disciplines (see Vandepitte & De Sutter (2013) for a survey) are all linked to this one element both fields of study have in common, i.e. "translations, which necessarily arise in the context of two different languages (or language varieties) and are therefore useful data types for both domains" (Vandepitte & De Sutter 2013, 36). Both the applicability of contrastive linguistic theories to TS as well as the acceptability of TS theory within contrastive studies are subject to debate. Whereas the use of translational corpora has received a rather straightforward acceptance in TS (see for example: Gellerstam 1986, 1996; Laviosa 2002), the debate about the inclusion of translational data within corpus-based contrastive linguistics is a more live one. The use of translations for contrastive research is indeed not without controversy and, seen from a TS point of view, the way in which translations are used in contrastive studies is often dismissed as unsuitable in a TS context.

With regard to the use of translations or translational corpora for contrastive studies, Altenberg and Granger (2002, 40) point out that the first attempt to compile a bidirectional electronic corpus for contrastive studies was made by Rudolf Filipovic and colleagues (Filipovic 1969). The researchers adopted the *translation method*, meaning that translators from the Yugoslav centers affiliated to the Serbo-Croatian and English corpus project were asked to translate parts of an existing corpus, *in casu* half of the Brown corpus (Filipovic 1969, 38-43)[8]. Despite the practical obstacles, contrastive linguistic researchers had indeed discovered the advantages of working with *parallel* corpora.

Apart from this early example of a parallel corpus, most bi- and multilingual corpora were only developed as from the 1990s (McEnery & Hardie 2012, 19) and within translation studies the so-called corpus turn coincided with the emergence of parallel corpora. Although McEnery and Hardie (2012, 20) claim that parallel corpora are typically used for translation research and comparable corpora for contrastive studies, we have already seen that this is only partially true. Comparable corpora can equally be (and have been) used for translation research (think of earlier mentioned research by Baker and Laviosa-Braithwaite) and parallel corpora have also been both extensively and fruitfully used in contrastive studies. Within contrastive studies, *translation equivalence* – necessarily established on the basis of *parallel* corpora – was considered "the best available *tertium comparationis*" (Johansson 1998, 5):

> Using the source or target language as a starting-point we can establish paradigms of correspondences (Johansson 1998, 5).

The usefulness of parallel corpora to establish *equivalence* was strengthened by the idea that source and target texts transferred "the same semantic content" (Granger 2003, 19). However, the assumption that translations could be used as a representation of "ordinary language use" – was as problematic for translation studies scholars as it was for contrastive linguists. In translation studies, this problem was countered by putting to the fore the investigation of translated language as a variety proper – thus clearly refuting the idea that translation could represent "ordinary language". In contrastive corpus studies, on the other hand, the idea arose that translations could be used as a *tertium comparationis*. One convincing argument as to why parallel corpora could be useful for contrastive linguists, is formulated by Noël:

[8] A second, smaller corpus was compiled consisting of a *few* Serbo-Croatian novels and their translations into English (translated by native speakers of English) (Filipovic 1969, 43).

> [T]he texts produced by translators can be treated as a collection of informants' judgments about the meaning of the linguistic forms in the source texts, with the added advantage that they are readily available to the linguist, who does not have to worry about constructing an experimental setup. Translation corpora can therefore be considered to be *a means of empirically testing one's intuitions* (or hypotheses) about the semantics of linguistic forms that is complementary to the systematic exploitation of the circumstantial evidence provided by monolingual corpora (Noël 2003, 759, our emphasis).

Aware of the fact that the results in TS were providing more and more evidence for the differences between translated and non-translated language, a number of scholars in contrastive studies did worry about what they called "translation effects" (Johansson 1998, 6) and proposed mechanisms to enable the researcher to control for those effects. One of those mechanisms is the procedure of *back-translation.*

2.3.2 Back-translation

Between 1969 and 1989, Vladimir Ivir published a number of articles (Ivir 1969, 1970, 1981, 1983, 1987, 1989) which were concerned with the notions of *formal correspondence* and *translation equivalence*, terms that had previously been coined by Catford (1965) from a translational perspective and by Ivir himself (1969, 1970) as well as a number of other scholars such as Krzeszowski (1971, 1972) from a contrastive linguistic perspective (Ivir 1981:51).

Ivir affirms that *"[f]ormal correspondence* is a term used in contrastive studies, while *translation equivalence* belongs to the metalanguage of translation" (1981, 51, our emphasis). Ivir is convinced that information from translations can be of valuable use to the contrastive linguistic analyst. His main concern is therefore to show "how translation equivalence enables the analyst to isolate formal correspondents" (Ivir 1981, 58). According to Ivir, *formal correspondents*, in the way defined by Catford "can hardly be said to exist" (Ibid., 54). He therefore proposes to adapt Catford's definition of *formal correspondence* so that it becomes defined "with reference to translationally equivalent texts" (Ibid., 55) rather than to linguistic systems. By re-defining *formal correspondence* in this way, it becomes a text-based, *equivalence*-based type of correspondence, in which the relationship between the correspondents is a one-to-many relationship (Ibid.) (one source language lexeme can yield many translation possibilities, and as a consequence, several correspondents). Ivir states that "formal elements which are correspondents in translationally equivalent texts [...] are matched in those of their meanings with which they participate in the particular source and target texts"(Ivir 1981, 55). He further on repeats that "such multiple correspondents are important analytical pointers to distinctions of meaning in the source language" (Ibid., 56). It is exactly this idea that will be exploited for the development of the Semantic Mirrors Method (see section 2.3.4) when using translations to lay bare different meanings. At all times, Ivir remains conscious about the difference in nature between translation (theory) and contrastive linguistic analysis (Ivir 1969, 15; Ivir 1970, 17; Ivir 1983, 173): translation aims at semantic *equivalence*s between texts, at the level of *parole* without the necessary need for consistent correspondence, while such formal-semantic correspondence is exactly the goal of a contrastive analysis at the level of *langue* (Ivir 1969, 15). While in nowadays (corpus-based) cognitive linguistics, the distinction between *parole* and *langue* has become somewhat obsolete – corpus-based cognitive linguistics is now conceived as "a usage-based approach to language that makes no principled distinction between language use and language structure" (Desagulier, 2014, 151) – the distinction was absolutely vital to contrastive linguists such as Ivir. His concern with the *langue* vs. *parole* dichotomy ultimately led to the formulation of a practical solution – *back-translation* – which allowed many corpus and contrastive linguists to fruitfully use translational data.

Ivir's main question with respect to translation is: "[h]ow much of the translated material produced by normal (unrestricted) translation can the contrastive analyst use?" (Ivir 1969, 16). In other words, how can the contrastive linguist detect or 'isolate' formal correspondents within translationally equivalent texts? (Ivir 1983, 175).

In answer to this question, Ivir proposes to apply the procedure of *back-translation* (first developed by Spalatin 1967), which preserves semantic content (Ivir 1987, 477) and relies on *translation equivalence* to isolate contrastive correspondents. The idea behind the *back-translation* procedure is the following: when an L2 item can be translated back into the (exact, same) original L1 item, no semantic shift takes place and the two items can be seen as contrastive (formal) correspondents. If, on the other hand, an L1 item different from the original L1 item is produced via *back-translation*, a "communicatively induced semantic shift" takes place and the two items cannot be regarded as contrastive correspondents (Ivir 1987, 477) unless the shift is due to "differences between the two linguistic systems" (Ivir 1983, 176). Next, a degree of overlap and difference between the L1 item and its paired L2 correspondents can be established by relating the L2 correspondents back to their expression in L1. Ivir remarks that, because of the L2 correspondents' polyfunctionality, each L2 correspondent will be related to a number of other L1 items too, besides the L1 with which the analysis was initiated (Ivir 1987, 478). The whole procedure of *back-translation* can be resumed in the following contrastive statements:

> When an L1 item has a given semantic function, its L2 correspondent is the L2 item A; for another function, its correspondent is the L2 item B, and for yet another the L2 item C, etc.; each of these L2 items, however, also corresponds to some other L1 items, resulting in a complex set of relations between the L1 item A and the L2 items A, B and C, then among the L2 items A, B, C, then between each of

> them and the L1 items A, B, C, D, E, F, G, and finally among the L1 items A, B, C, D, E, F, G. Conditions can be specified for these relations, which, together with the listing of multiple correspondences, are exploited in pedagogical and other applications of contrastive analysis (Ivir 1987, 478-479).

Schematically, the procedure then looks as follows (adapted from Ivir 1987, 478):

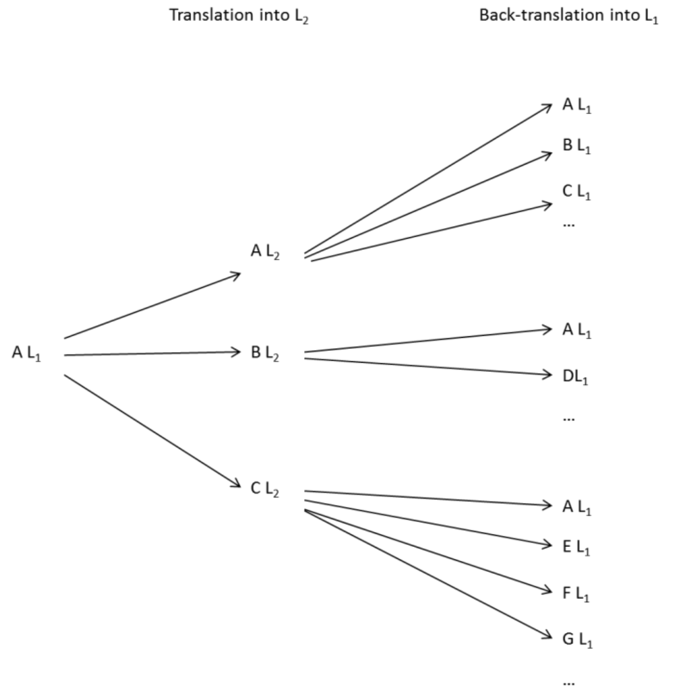


Figure 2 Back-translation procedure for contrastive analysis (Ivir 1987, 478)

To resume, *back-translation* was initially developed by Ivir as a contrastive-linguistic tool or procedure to identify *formal correspondents* (redefined by Ivir as contrastive correspondents) within translational data, therefore relying on a usage-based relation of *translation equivalence*.

Two additional advantages of the technique need to be pointed out here. First, the procedure of *back-translation* enables the researcher to lay bare the one-to-many relationship between an L1 item under scrutiny and its L2 contrastive correspondents and can therefore possibly lay bare meaning differences:

> The relationship between an L1 unit and its L2 correspondents is not one-on-one but one-to-many, with each L2 correspondent matching a particular segment of the meaning of the L1 unit but also introducing other meanings which the L2 units has in the set of oppositions in that language (Ivir 1983, 177).

Second, Ivir's concern with the distinction of *contrastive correspondents* equally allows the (translation studies) researcher to separate "irrelevant differences that are due to the translator's idiosyncrasies or motivated by particular communicative or textual strategies" (Altenberg & Granger 2002, 17) from – what Dyvik will call – Linguistically

Predictable Translations (see p.48). All this points towards the suitability of the *back-translation* procedure for (contrastive) research into meaning (based on translational data) as well as for (corpus-based) investigations of meaning in translation.

2.3.3 Applying back-translation: Mutual Correspondence

The idea of *back-translation* has been further used and developed within contrastive linguistics. As was already mentioned in the introduction of this section, the use of translational corpora has received a rather straightforward acceptance in TS (see for example: Gellerstam 1986, 1996; Laviosa 2002), compared to contrastive studies. The mentioned consensus about the use of translation corpora as "an empirical basis for semantic claims" (Noël 2003, 758) received further support from Ebeling & Ebeling (2013, 24-28), who consider that a supporting basis of the use of translational data and parallel corpora for contrastive analysis can in fact be found either in Ivir's work on *back-translation* or in Altenberg (1999) and Altenberg & Granger's (2002) work.

Altenberg and Granger's proposition builds on the idea of *back-translation*, and does as such not provide a distinct line of argumentation. Their application is called Mutual Correspondence (Altenberg 1999, 254 ff.; 2007, 9; Altenberg & Granger 2002, 18) (hence: MC) and combines the idea of *back-translation* with a quantitative *equivalence* concept (such as Krzeszowski's notion of *statistical equivalence* (1990, 27-28)) in order to obtain more evidence about the relevance of the detected translation patterns (Altenberg & Granger 2002, 17):

> 'Mutual correspondence' (MC) is a simple statistical measure of the frequency with which a pair of items from two languages are translated into each other in a bi-directional translation corpus (see Altenberg 1999). This can be calculated and expressed as a percentage by means of the formula:
>
> $$\frac{\mathrm{A_t} + \ \mathrm{B_t}}{\mathrm{A_s} + \mathrm{B_s}} \times 100$$
>
> where At and Bt are the frequencies of the compared items in the translations, and As and Bs their frequencies in the source texts. The value will range from 0 (no correspondence) to 100 (full correspondence) (Altenberg 2007, 9).

MC exploits Ivir's notion of *formal correspondence* – established via *back-translation* – while adding a quantitative aspect to it. Gilquin (2008) praises the possibility *back-translation* offers "to control for translation effects ("translationese", cf. Gellerstam 1986) by taking into account the " "inverted" equivalence" (Gilquin 2008, 186) and uses MC as a cross-linguistic measure of *equivalence* between two words or constructions (Ibid.). Mortier (2010) describes her use of MC as the establishment of "the degree to which source and target items correspond in the two languages" (Mortier 2010, 410). Both applications agreeably emphasize that MC is a contrastive measure which holds between different language items, not between same language items: one can only calculate an MC between an L1 item *a* and an L2 item *z*, or between an L1 item *b* and an L2 item *y*, but MC does not provide the researcher with any (direct) information about the monolingual relationship between the two L1 items *a* and *b*. Furthermore, the resultant correspondences are calculated for each of the contrastive pairs individually; the overall 'network' of relationships between the source language lexeme(s) and all attested translations stays somewhat out of the picture.

Although MC appears to be an interesting application of *back-translation* for semantic research, it is, due to its clear contrastive nature, incompatible with our objective to compare semantic field representations in translated and non-translated language, a comparison which involves different representations of one language.

2.3.4 Applying back-translation: Semantic Mirroring

A second application of *back-translation* can be found within automatic thesaurus extraction. The semantic mirrors method was first introduced in 1998 as a solution for automatic thesaurus building and underwent further development within the project "From Parallel corpus to Wordnet" which was carried out at the University of Bergen (2001-2004) (Dyvik 2004, 311). The project explores the use of translational data as a basis for semantic research. Possible applications of the technique are the derivation of "large-scale semantically classified vocabularies" for machine translation and other types of multilingual processing (Dyvik 1998, 51) and later also the derivation of wordnet relations within the previously mentioned project (2004, 311)[9]. The idea of the SMM – which will be at the heart of the methodological tool we aim to develop – in fact finds itself at this crossroads of linguistic software development and lexical-semantic investigations. In this section, we will explore how the SMM can be a possible answer to the investigation of meaning relationships in translation.

2.3.4.1 Selecting translational data

First, and in an effort to hold the balance between computational linguistic pragmatic solutions and a traditional lexical semantic reticence to use translational data, Dyvik (i) puts forward a number of strong arguments in favor of translation and (ii) focuses on what he calls the *translational relation*, a notion that will underpin his translation-driven technique for lexical semantic investigation, i.e. the SMM.

[9] We will discuss applications of the SMM by other researchers and outside these two fields of application in section 2.3.4.4. Obviously, our own 'extension' of the method, which will be presented in chapter 3 is also a kind of application.

According to Dyvik, the semanticist first needs to get persuaded of the usefulness of translation for linguistic semantics. Apart from the fact that translation is a large scale activity, bringing about a multi-lingual perspective on lexical semantics, Dyvik additionally and convincingly argues that translation is *an evaluation of meaning in a normal kind of linguistic activity*, outside any kind of metalinguistic, philosophical or theoretical reflection (Dyvik 1998, 51). It can therefore provide the researcher with strong empirical evidence, viz. the relations between the texts which are the observable results of the translator's evaluation of the meaning under scrutiny (Ibid.).

Exactly because translation is such a normal, omnipresent type of activity, the *translational relation* can be said to emerge "as epistemologically prior to more abstract and theory-bound notions such as 'meaning', 'synonymy' and 'inference' " (Dyvik 2005, 27). This assumption suggests that the *translational relation* between languages can be taken as *a theoretical primitive,* "a concept not to be defined in terms of other concepts, but assumed to be extractable from translational data by interpretive methods" (Dyvik 2005, 27). Following Dyvik, we accept that the translational relation can indeed be 'extracted' from translational data. It is furthermore the impossibility of "a perfect translation" which makes translation so interesting for the semanticist:

> Languages [...] are discrete structures, and meanings are entwined in the structures themselves. Therefore, during translation, things crack and snap, things disappear, and things are added, and there is hardly ever a unique correct solution to a translational task. Instead, actual translations provide a host of alternative approximations to the unattainable ideal, and this is a potential source of information: semantic insights may emerge from the way the sets of alternatives are structured (Dyvik 2005, 28).

2.3.4.2 Translationally derived features

Convinced about the *acceptability* of the use of translational data, Dyvik's first concern when working with this type of data is to select the *adequate* data (1998, 52): the contribution of contextual factors should be separated from the correspondence relations, the latter being the type of relations the (contrastive) semanticist is interested in. For this reason, (translational, parallel) corpus data cannot be used in their raw form: so-called 'bad translations' need to be filtered out of the data, and Linguistically Predictable Translations[10] need to be isolated from the totality of the data (Ibid.), and

[10] A Linguistically Predictable Translation is a translation that is not (completely) dependent on "the particular text and its circumstances" (Dyvik 1998). E.g. the translation of Dutch *huis* in the source language sentence *hij woont in een mooi huis* [he lives in a beautiful house] by English *house* in the target language sentence *he lives in a beautiful house* is linguistically predictable. On the other hand, the translation of *huis* in the sentence *ieder huisje heeft zijn kruisje* [every house has its crucifix] by *cupboard* in the target language

consequently selected for further analysis. Dyvik's decision to select LPTs is driven by the same concern as Ivir's selection of *contrastive correspondents* ("how much of the translated material can the contrastive analyst use"). Dyvik arrives at the selection of the LPTs by applying a procedure which is very similar to that of Ivir's *back-translation*[11]. The difference between Ivir's and Dyvik's proposal lies in the purpose for which they apply *back-translation*: where Ivir's sole concern is to select contrastive pairs, Dyvik moreover aims to *generate* 'new' semantically informative information (about synonymy, hyponymy, etc. to suit his thesaurus building purposes), and he does so by applying the method to a parallel corpus.

The semantic informativity of the procedure can be understood as follows. Consider, for example[12], the Dutch noun *heks,* which can be translated into English as *hag* and *witch.* According to Dyvik, the fact that alternative translations exist, points towards a relatedness to either different 'aspects' or different sub-senses of the meaning of *heks*: the English words indicate one of the many possible ways of dividing the semantic potentiality of *heks* (Dyvik 2005, 31).

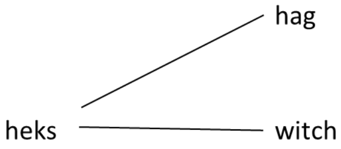


Figure 3 Translational correspondence

Subsequently, the lexical sub-senses of *heks* could be expressed as contrastive pairs: <heks, hag>, and <heks, witch>. Within a translational approach, these *pairs* (called *sets* when several languages are involved) can be seen "as a kind of *semantic features*, [...] assignable to lexical items, both to the items they were derived from, and to others, which may inherit them [...]" (Dyvik 2005, 31, our emphasis). Schematically, the "translationally derived features" would then look as follows:

sentence *there's a skeleton in every cupboard* is not linguistically predictable because it depends on the particular context, in this case, the idiomatic expression in which it is used.

[11] We will see in chapter 3 that our proposal for an extended SMM is also based on a procedure which includes the use of a – be it differently operationalized – type of *back-translation.*

[12] This example is adapted to the Dutch-English language pair from Dyvik's (2005, 29-31) German-English example.

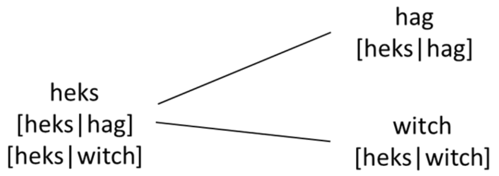


Figure 4 Translationally derived features

To sum up, semantic information can be obtained from *translationally derived features*:

> Intuitively, the features encode subsenses that the lexical items share with each other. In this way the features become *classificatory devices*, grouping lexical items together according to shared semantic properties (Dyvik 2005, 31, our emphasis).

In a classical structuralist approach, the semanticist would describe word meaning via a *componential analysis*, in which he assigns *semantic features* to words, in order to understand their interrelations (Dyvik 2005, 28). While it is true that from a purely structuralist point of view translations could never be used as contrastive semantic informants – because different languages carve up the world or a same semantic field in different ways – Dyvik observes that these differences in carving up the same field are reflected "in the fact that this translational relation is not one-to-one" (Ibid., 29) and are semantically informative: contrastive differences can be a reflection of difference(s) (in classification) of semantic properties. Dyvik explicitly states that meaning can be inferred from the *translational relation* between a source language (lexeme/structure) and its translation:

> Corresponding sets of terms in two languages are connected by a relation of translation (Dyvik 2005, 29).

> The translational relation between the signs of two languages (interrelating 'linguistically predictable translations') is an instance of the sharing of meaning properties across languages (Dyvik 1999, 217).

In other words: a *translational relation* cannot exist between an LPT and its source language lexeme if these two do not share any meaning properties (Dyvik 1999, 218). Translational properties can be 'easily' accessed – at least more easily than the much more abstract meaning properties – by investigating source texts and their translations. It can therefore be tried to "define (some) meaning properties in terms of translational properties rather than the other way around (as is common)" (Dyvik 1999, 218). In Dyvik's view – which we will adopt for our extension of the SMM (see section 3.4) – *semantic features can be derived from translational data*: alternative translations are related to different aspects or related sub-senses of the meaning of a word under scrutiny (Dyvik 2005, 31), and can divide up the semantic potentiality of the given word (Ibid.). In this way, "sets of translationally corresponding items across languages [can be seen] as *the primitives of semantic descriptions*" (Ibid.), and the contrastive pairs can be considered as semantic features, assignable to lexical items (Ibid.).

2.3.4.3 Ivir and Dyvik

Throughout our explication of Dyvik's ideas, we have seen that there are quite some similarities with Ivir's ideas about *contrastive correspondents* and *back-translation*, although Dyvik seems to develop his ideas independently of Ivir's previously established notions. Dyvik's and Ivir's proposals are similar in that they (i) each use a mechanism which allows them to select only those translational data which they find suitable and 'safe' for contrastive analysis; and (ii) treat the relation of *translational correspondence* as a symmetric relation "disregarding the direction of translation" (Dyvik 2004, 314), a viewpoint which is in line with their research goal and seems for both Ivir and Dyvik the methodologically right thing to do: in their contrastive view, pairs of translations are informative tools used for their dynamics to 'move' between languages in a meaning-preserving way, informing the researcher about meaning, while the influence of the task of translation itself is brought down to a minimum, so that the data are as 'contrastively pure' as possible. From a point of view of translation studies though, the *translational relation* is clearly asymmetric and this has been proven via the same practice of *back-translation*: "[m]ultiple examples from the practice of back translation have proven that translation pairs are not symmetric and translation through several languages make the lack of transitivity similarly apparent (see e.g. Levý 1989)" (Halverson 1997, 211). Within our use of this method, we will necessarily have to take into account the asymmetry of the *translational relation*, since we are interested in translation itself, and not merely in its exploitation as a (logical) tool.

One could wonder why we are making such an effort to present Dyvik's technique, if in fact Ivir's previously formulated ideas were so similar. There are several important reasons for us to prefer the SMM as a basis for our methodological tool to 'pure' *back-translation* as formulated by Ivir. First, Dyvik makes an important link between a technique, *back-translation* and a specific research objective: lexical semantic research, an objective which we share with Dyvik. As a matter of fact, Dyvik operationalizes Ivir's intuition that each L_2 correspondent will be related to a number of other L_1 items too, besides the L_1 with which the analysis was initiated (Ivir 1987, 478) by retrieving the "other L_1 items" in an additional corpus-based retrieval step (called the *inverse T-image*). Second, as Ebeling & Ebeling (2013, 25) rightly remark, Ivir never explains the procedure of *back-translation* in detail, which makes it difficult to know whether he applies the method with a parallel corpus or if *back-translation* is done on the basis of the analyst's translational intuitions. For Dyvik on the contrary, the use of (parallel) corpora is an obviousness, explicitly mentioned in his design. Again, we share Dyvik's view to explicitly put forward a parallel corpus approach for research in lexical semantics of translation. Finally, Dyvik further develops and exploits a notion (which was also mentioned by Ivir, but not exploited) i.e. *overlap* to ensure the semantic relatedness between the yielded lexemes, and this notion is part of a procedure of *back-and-forth translation*, an additional dimension which will be exploited for the comparison of translated and non-translated language (see section 3.4 on extended semantic mirrors).

2.3.4.4 The SMM in contrastive linguistic studies

Within contrastive, corpus-based studies, Aijmer and Simon-Vandenbergen have drawn extensively on Dyvik's idea of using translations as 'mirrors' for semantic field research. They mainly focused on discourse particles (Aijmer & Simon-Vandenbergen 2003), pragmatic markers (Simon-Vandenbergen & Aijmer 2002-2003; Aijmer & Simon-Vandenbergen 2004; Aijmer et al. 2006) and adverbs (Simon-Vandenbergen & Aijmer 2007; Simon-Vandenbergen 2013)[13]. In line with the cautiousness which contrastive researchers usually show when employing translational data, Aijmer and Simon-Vandenbergen relied on Dyvik's argumentation to legitimately incorporate into their analysis the supplementary information which translations are able to provide about semantic similarity. They show an interest in using the *back-and-forth translations* as a *tertium comparationis* (Simon-Vandenbergen & Aijmer 2002-2003, 16; Aijmer & Simon-Vandenbergen 2004, 1795), but their main interest in Dyvik's proposal stems without a doubt from its aptitude to construct and compare semantic fields (Aijmer & Simon-Vandenbergen 2003, 1131; 2004, 1782; Simon-Vandenbergen & Aijmer 2002-2003, 13), a goal that we share with both Aijmer & Simon-Vandenbergen and Dyvik. A meticulous summing up of the differences between Dyvik's method and the way in which Aijmer and Simon-Vandenbergen put it into practice would be of little use to the development of our own application of the SMM, which will necessarily have to fit the peculiarities of our own research objectives. We will therefore single out a number of adaptations and specifications made by Aijmer and Simon-Vandenbergen, which will be of interest to our understanding and adaptation of the method.

1. Aijmer and Simon-Vandenbergen always use at least three languages; i.e. the language under study (English) and two mirror languages: either Dutch and

[13] Mortier & Degand (2009) were inspired by the work of Aijmer and Simon-Vandenbergen and carried out a "mirror-analysis" for adversative discourse markers. Mortier & Degand combine different types of corpora (parallel and comparable, with written and spoken data) to arrive at a "semantic profile" for the discourse markers under study. They emphasize that their application of the "mirror analysis" serves to establish "the field of formal equivalents in one language or across languages" (Mortier & Degand 2009, 309). According to the researchers, a mirror analysis "consists of back-and-forth translations of a given item from the source language to the target language, and form the target language back to the source language". Their application of the procedure in fact answers perfectly to Ivir's *back-translation* procedure for the retrieval of *formal correspondents* (and this is also the goal of Mortier and Degand), so their method stands much closer to Ivir's contrastive notion than to Dyvik's lexical-semantic tool.

Swedish (Aijmer & Simon-Vandenbergen 2004), or Dutch and French (Simon-Vandenbergen 2013) or even four mirror languages (Dutch, Swedish, French and German) at once (Simon-Vandenbergen & Aijmer 2007), whereas Dyvik uses two languages: one language under scrutiny and one pivot language. Aijmer and Simon-Vandenbergen in fact combine the resulting translations from two mirror analyses (a mirror exercise can only be carried out with one language at a time) into one resultant relational field. If, for instance, Dutch and Swedish are used as pivot languages, this double mirror allows them to compare the *overlapping translations back into English*. *Overlapping translations* are interpreted here as those translations back into English which are obtained as translations of both Swedish and Dutch source lexeme(s). The result is a set of English lexemes, *overlapping*[14] between Dutch and Swedish. Aijmer and Simon-Vandenbergen compare in this way the number of identical translations (from Dutch or Swedish) into English yielded in what they call "the second translation image" (Aijmer & Simon-Vandenbergen 2004, 1796), which corresponds to Dyvik's step of the *inverse T-image* (see section 3.3.1). Combining different mirror images into one result, also implies that data are obtained from different corpora and need to be combined while staying comparable.

2. Whereas Dyvik's "ranking of signs in a semantic field" is done "quite independently of frequency of occurrence"[15] and based on the "overlap relations among *t*-images" (Dyvik 1998, 73)[16], Aijmer and Simon-Vandenbergen (2004) use frequency information to differentiate the items of a lexical set (obtained via a mirror analysis as translations of one particular marker in one language under scrutiny):

> Such paradigms or lexical sets show, for example, which translations are more frequent or prototypical, and which are less frequent or even 'singleton' translations (Aijmer & Simon-Vandenbergen 2004, 1785-1786).

The (relative) frequency information of correspondences is used to distinguish between prototypical equivalents and more context-bound correspondences

[14] Note that this interpretation of overlap differs from our interpretation of the notion (see section 3.3.3).

[15] "(except that a lexeme of course has to occur at least 32 times in the corpus in order to be a member of 32 subsets)" (Dyvik 1998, 73).

[16] Recall the quote at the beginning of this section, stating that overlapping first *t*-images do not guarantee that two lexemes indeed pertain to the same field "since the shared L2 sign may be ambiguous between an '*a*-sense' and a '*b*-sense' with no close relationship between them" (Dyvik 1998, 72). In order to ensure that two lexemes do pertain to the same field, Dyvik proposed the technique of *back-and-forth* translation up to the level of the *second T-image* (the necessity of the *second T-image* will be further explicated in the methodological chapter of this study).

(Simon-Vandenbergen & Aijmer 2007, 8), but frequency information is not as such integrated in the visualized results which represent the *translation networks* (Simon-Vandenbergen & Aijmer 2007, 250-253). The researchers choose to only consider salient correspondences in their translation network "in principle the five most frequent ones, though individual decisions had to be taken in view of the large differences in absolute and relative frequencies in separate tables" (Ibid., 248). This problem is a direct consequence of the fact that different corpora had to be combined for this application. Conclusively, Aijmer and Simon-Vandenbergen do not neglect frequency information, but the resultant contrastive *translation networks* are not (directly) based on the frequencies of the correspondences; the lines which link up the contrastive lexemes in the *translation networks* in fact only reflect cross-linguistic translation overlap[17], which is a different kind of overlap from Dyvik's notion. A distinction is made between full lines to mark the prototypical correspondences, and broken lines which show "correspondences which are not prototypical but [...] still recurrent enough to be included" (Ibid., 248).

3. Aijmer et al. (2006) explicitate a step in their use of translational data which will be of particular importance to our use of the technique:

> In order to conclude that two items belong to the same semantic field, it is not sufficient to look at translations in one direction only; one must *go back and forth* from sources to targets. If item X in language A is translated by Y and Z in language B, one can, by using B as source language, look for translation equivalents of Y and Z (see also Ebeling 2000, 17-18) (Aijmer et al. 2006, 112, our emphasis).

Until now, the expression *going back-and-forth between sources and translations* has itself remained ambiguous: linking translations and sources can indeed be done within a same set of data, but if one only looks back at the source language lexemes of the translations, the sole intention can be to disambiguate the given source language lexeme(s). If we wish to create a semantic field, and engage in the selection of a number of candidate-lexemes for this particular semantic field, then the used technique will necessarily be expansive, meaning that it will have to yield new information, i.e. new candidate-lexemes. To arrive at this goal, the solution is here to do as Dyvik's method in fact implies and as Aijmer et al. (2006) make more explicit: if, in the step of the *inverse T-image* (see section 3.3), language B is used as a source language, then new translations

[17] This *modus operandi* is further confirmed in Simon-Vandenbergen (2013, 93-94), where the relation (within a 'mirror analysis') between French or Dutch equivalents and English lexemes is indicated by one cross if such a relation exists and two crosses is the relation was recorded more than once.

(i.e. language A lexemes) which are semantically related to the *initial lexeme* are revealed and can be considered candidate-lexemes for the intended semantic field.

To sum up, Aijmer and Simon-Vandenbergen propose a "translation-based variant of semantics based on data from translation corpora" (Simon-Vandenbergen & Aijmer 2007, 7) for which they draw on Dyvik's semantic mirrors method. Interesting adjustments to the technique consist in (i) their use of multiple languages to arrive at a final semantic map, (ii) the integration of frequency information, although without statistically incorporating this information into the analysis and (iii) the explicitation of the data-extraction procedure. Note that Aijmer and Simon-Vandenbergen do not explicitly refer to the different steps of the SMM as referred to by Dyvik, but from the given description, it can be deduced that they do not apply the last step of the SMM, i.e. the *second T-image.*

2.3.4.5 The SMM in other domains of linguistics

The SMM has also drawn the attention of researchers in Natural Language Processing. Priss & Old (2005) have proposed to model the SMM with Formal Concept Analysis, using concept lattices instead of the Venn diagrams proposed by Dyvik to visualize semantic relatedness. Eldén et al. (2013) propose to visualize the semantic relations which come from semantic mirrors via Spectral Graph Partitioning. In addition to this, the SMM has been compared, within the realm of computational linguistics, with its 'competing' distributional techniques for automated thesaurus construction. Muller & Langlais (2011) concluded that "with respect to synonyms, [...] mirror translations provide a better filter than syntactic distribution similarity" (p.333). It is beyond the scope of this study to further comment on these computational applications, but the fact that the SMM has been applied both in more theoretical contrastive linguistic works on the one hand and in computational applications on the other at least shows that the ideas underlying the SMM have found support in both theory and in practice.

2.3.5 Conclusion

In this section, we have shown how *back-translation* can be used as a contrastive linguistic tool, able to isolate *formal correspondents* (renamed and re-defined by Ivir as contrastive correspondents) and capable of detecting semantic relationships between lexemes in one language. An application of *back-translation* via *semantic mirroring* offers – in theory – the possibility to investigate semantic relationships in translated and non-translated language because it makes use of parallel corpora. Although the SMM has indeed the *potential* to lay bare meaning relationships (the application of *back-and-forth translation* as well as *overlap* will lead to the selection of a set of lexemes which are semantically related to each other, see section 3.4.1), a number of issues remain unsolved. First, we are still lacking a notion of *translation equivalence* that is operationalizable (within the procedure of *back-and-forth translation*) in such a way that valid comparisons between translated to non-translated language can be made. As we have laid out in this chapter, both Dyvik and Ivir establish *equivalence* on the basis of a symmetric notion of the translation relation, but the idea that *equivalence* is symmetric is incompatible with our CBTS viewpoint. Second, the SMM was originally a method for thesaurus building and is therefore not 'equipped' to carry out comparisons of the semantic relationships it lays bare amongst different language varieties. Thirdly, provided that we overcome the first two issues, we are still missing a theoretical framework within which we can interpret those comparisons. Solutions to each of these problems can be found within corpus-based semantics.

2.4 Corpus semantics

In this section, we will provide theoretical insights from different areas of corpus-based semantics. Before we can pursue to establish our methodology, three elements are still missing: (i) an acceptable notion of *translation equivalence* (applicable within the SMM and allowing an asymmetric translational relation), (ii) an *insightful* means to straightforwardly compare semantic relationships in translated and non-translated language and (iii) a theoretical framework within which such comparisons can be interpreted. Corpus(-based) semantics is an extremely vast area of research. We will therefore only touch upon those domains that are immediately relevant to theoretically underpin the three aspects cited above.

In the first part of this section (2.4.1), we deal with the notion of *translational equivalence* as it was developed in Word Sense Disambiguation. We will see that, by considering *translational equivalence* according to its WSD-based definition, the notion can also be used when the translational relation is not considered symmetric (as is the case in our study).

In section 2.4.2, we will see that the semantic relationships revealed on the basis of the *translational equivalence* hypothesis can be understood in terms of distances and captured in so-called Semantic Vector Spaces. Statistical visualization methods can consequently be used as "an intuitive interface" (Heylen et al. 2012, 17) to study semantic relationships in fields of translated and non-translated language.

In section 2.4.3, we will explore how the idea of a "prototype model of category structure" – considered as one of the important contributions of cognitive semantics to the study of word meaning (Geeraerts 2013, 577) – can form the theoretical background against which the semantic relationships within the semantic field under study can be interpreted.

2.4.1 Translational equivalence in Word Sense Disambiguation

The idea that a procedure such as *back-translation* based on *translation equivalence* introduced in section 2.3 can be used to lay bare semantic relationships also exists within corpus-based semantics. The derivation of semantic relationships on the basis of *translational equivalence* is put into practice within Word Sense Disambiguation – a name commonly given in the field of computational linguistics to the task of "computationally determining which "sense" of a word is activated by the use of the word in a particular context" (Agirre & Edmonds 2007, 1).

In WSD, *unsupervised corpus-based methods*[18] are either based on the distributional hypothesis, or, alternatively, on the idea of *translational equivalence* (Agirre & Edmonds 2007). So-called distributionalist methods are often summarized in John R. Firth's well known words "You shall know a word by the company it keeps" (Firth 1957, 11) [19]. The *translation equivalence* hypothesis is based on the idea that a word can be known by the *translational* company it keeps. *Translational equivalence* methods were introduced into computational linguistics because of their relevance for machine translation (Pedersen 2007, 134), one of the earliest fields of application of WSD. The reliability of *translational equivalence* has received direct evidence from WSD: according to Ide, Erjavec and Tufiş (2001, 1) "sense distinctions derived from cross-lingual information correspond to those made by human annotators, especially at the coarse grained level" and "the reliability of sense assignments at finer-grained levels is comparable for human annotators and those produced automatically with cross-lingual data".

While in lexical semantics, distributional approaches are widely applied[20], methods that rely on *translational equivalence* as a meaning-structuring device have not yet had

[18] The different approaches to WSD are classified according to their main source of information: *knowledge-based methods* use sources such as dictionaries and thesauri, *unsupervised methods* collect information from raw unannotated corpora and include methods using word-aligned corpora which extract cross-linguistic information; *(semi-)supervised methods* train from annotated corpora, or use them to seed in a bootstrapping process (Agirre & Edmonds 2007, 12).

[19] In computational linguistics, the distributional hypothesis is also commonly attributed to Wittgenstein (1953), Harris (1954) or Weaver (1955) (Turney & Pantel 2010, 142-143).

[20] In lexical semantics and lexical variation studies (e.g. Peirsman et al. 2010), the distributionalist idea has led to the advent of (semi-)automatic retrieval methods of semantically similar words such as latent semantic analysis (Landauer & Dumais 1997) first and second order bag-of-words models (Manning & Schütze 1999) and the behavioral profiles method (Divjak & Gries 2006, 2009).

much uptake. 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 fact, the idea of *translational equivalence* can be rather straightforwardly related to the widely used distributional approach. We could easily reformulate the acceptability of *translational equivalence* in distributionalist terms, i.e. with respect to the (additional or alternative) contextual disambiguation possibilities that translations offer: the addition of information from a second language (a translation) about a lexeme under scrutiny (the source language lexeme) – which stands in a translational relation to that lexeme – can be seen as 'addition of context'. *Translational equivalence* methods could therefore be said to form – at least conceptually, and at least for research focusing on lexical semantic investigations in translation studies – a possible alternative for or addition to the existing distributional methods, as is already the case within WSD.

How does translational equivalence work in WSD?

Now that we have argued in favor of the conceptual acceptability of *translational equivalence* for lexical semantic research in translation, we need to understand exactly how *translational equivalence* works within WSD.

WSD methods based on *translational equivalence* unsurprisingly use *translations* as information source for disambiguation:

> methods based on *translational equivalence* rely on the fact that the different senses of a word in a source language may translate to completely different words in a target languages (Pedersen 2007, 134)

In machine translation (the field where WSD researchers initially got the idea for *translational equivalence*), "the ambiguity of a source word is [...] given by the number of target representations for that word in the bilingual lexicon of the translation system" (Dagan et al. 1991, 132). For example, if in a machine translation task, the correct sense of the English lexeme *bank* needs to be selected, the *conditio sine qua non* to perform this task (correctly) is that the system disposes of the necessary information to differentiate between the different senses of *bank.* The distinctive senses of *bank* can be assigned to the lexeme "by producing all the [French] alternatives for the lexical relations involving [*bank*]" (Dagan et al. 1991, 131). The French translation *banque* distinguishes the "financial institution" sense of *bank*, whereas the French *rive* reveals the 'riverside' sense of *bank.* Schematically, the sense assignment looks as follows:

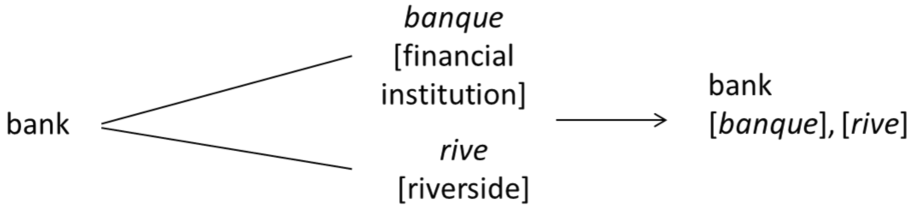


Figure 5 Different senses of the English lexeme *bank* are assigned based on its French translations

Given that the lexeme *bank* now disposes of two possible senses, it has become possible to select the sense "which corresponds to the most plausible [French] lexical relations" (Dagan et al. 1991, 131) and consequently to select the contextually correct target word.

Not all ambiguities can be resolved through 'simple' *translational equivalence*. For instance, at least two senses of the Dutch lexeme *school* cannot be disambiguated while using English translations: the "educational institution" sense of Dutch *school* translates in English as *school,* and also the "group of fishes" sense of Dutch *school* translates into English as *school,* and ambiguity remains unresolved (Figure 6). In these cases, it is proposed to add a third language (Dagan et al. 1991, 132). In this particular case, adding French would help, as the 'group of fishes' sense translates in French as *banc,* and would reveal this additional sense (Figure 7).

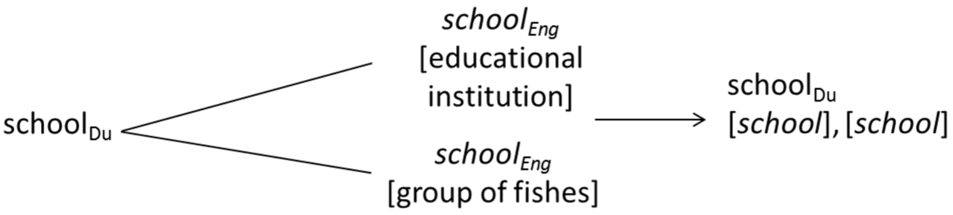


Figure 6 Unresolved disambiguation via one language

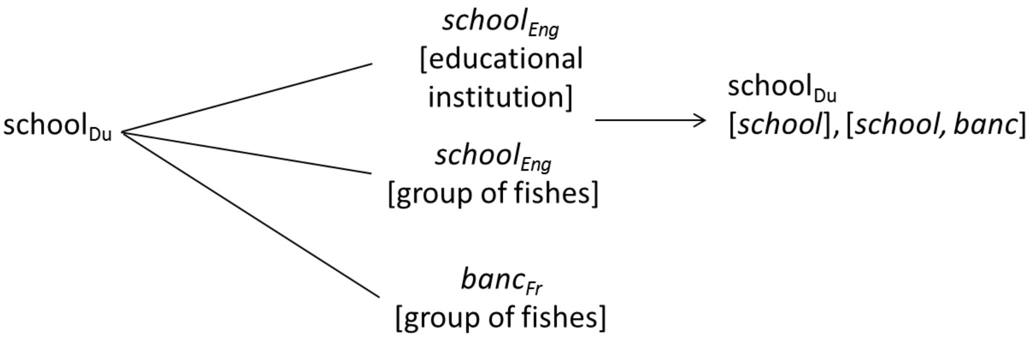


Figure 7 Resolved disambiguation via two languages

While adding a language or even several languages (Lefever et al. 2013), has proven to be an effective way to enhance the WSD procedure, it is also conceptually possible to rely on a single language and still arrive at the disambiguation of the different senses. This can be done by applying the procedure of *back-and-forth translation* following the

SMM. Within the SMM, the translational relation is, however, considered as symmetric, an idea which is incompatible with our translational point of view that translation is necessarily asymmetric (see section 3.4.1). The idea of a symmetric *translation equivalence* relation is, however, not a prerequisite to carry out *back-and-forth translation* with the SMM. In fact, disambiguation via the SMM can rely on the same basic idea as disambiguation via several languages in WSD, which states that the different senses of a word are determined by considering only those distinctions that are lexicalized cross-linguistically (Ide & Wilks 2007, 54) – no more, no less. By considering the relation of *translational equivalence* in the SMM as identical to the one in a WSD disambiguation task with several languages – not necessarily symmetric and lexicalized cross-linguistically – we can use the SMM for a disambiguation task within this study which aims to investigate semantic differences between translated and non-translated language.

2.4.2 Vector Space Models

In the previous sections, we have seen that although the SMM can be used to reveal semantic relationships, it cannot be readily used to compare the obtained relationships amongst different language varieties. The same holds for WSD: it is a (computational) task to determine sense distinctions, but does not offer solutions as to how to objectively compare the disambiguated senses. Objective comparisons would indeed require objective visualization methods, which neither the SMM nor WSD straightforwardly offer. In this section, we will turn to linguistic semantics and corpus-based cognitive semantics, which are mainly occupied with the empirical study of lexical meaning. We will see that the semantic relationships revealed on the basis of the *translational equivalence* hypothesis can be understood in terms of distances and captured in so-called Semantic Vector Spaces (SVS). Statistical visualization methods can consequently be used as “an intuitive interface” (Heylen et al. 2012, 17) to explore the semantic relationships in fields of translated and non-translated language “captured by an SVS” (Ibid.).

In linguistic semantics and corpus-based cognitive semantics, the perceived difficulties to introspectively analyze meaning and meaning differences have led to the development of “a methodology for empirical research in cognitive linguistics that is based on thorough quantitative analysis of corpus data” (Heylen et al. 2008, 91). Data are derived from or gathered via corpora and quantitatively analyzed using methods that are “methodologically similar” to work in computational linguistics or information retrieval (Gries 2006a, 6). Geeraerts (2016, 242) and Stefanowitsch (2010) discern three major perspectives: experimental research, the referential method and the distributional, corpus-based approach. As we have seen in section 2.4.1, our own proposition to reveal semantic relationships on the basis of the translational hypothesis can be fitted in with the distributional, corpus-based approaches to the empirical study of lexical meaning as translations can be considered as an alternative for or additional type of context.

The distributionalist corpus-based method takes three main forms (Geeraerts 2016, 242-243): one in the tradition of Sinclair, a second one following the *behavioral profile* approach and a third form, called the *semantic vector space* approach. In Sinclair's tradition, statistical methods are used to "identify semantically relevant contextual clues in the corpus" (Geeraerts 2016, 242) after which the "semantic characterization" of the words and expressions is usually analyzed manually (Geeraerts, Ibid.). The *behavioral profile* approach takes the opposite direction: potentially interesting features are first tagged manually or semi-automatically, after which statistical techniques are applied to "classify the occurrences into distinctive senses and usages" (Geeraerts 2016, 243). Various statistical techniques have been used within this approach, e.g. hierarchical cluster analysis by Gries (2006b) and Divjak (2010) and correspondence analysis by Glynn (2010) (see Geeraerts 2016, 243). The third approach, the *semantic vector space* approach uses quantitative techniques on both levels: contextual clues are first identified in a statistical way; the subsequent "clustering of occurrences on the basis of those clues" is equally carried out statistically (Geeraerts 2016, 243).

Vector Space Models (hence: VSMs) – which are put forward within this *semantic vector space* approaches – were initially proposed as a solution to the problem of document retrieval in Information Retrieval (Clark 2015, 495). They can be combined with the distributional hypothesis "as an approach to representing some aspects of natural language semantics" (Turney & Pantel 2010, 141). Ruette et al. (2014, 212) explain how VSMs can be combined with the distributional hypothesis:

> [I]n Vector Space Models, objects are described by *n* quantifiable characteristics. These characteristics make up an *n*-dimensional space in which the objects can be positioned. Every characteristic is thus a dimension. The position of the objects along these dimensions depends on the value that the characteristics have. In a way, these values can be seen as coordinates of a point in the *n*-dimensional space, made up by the characteristics. The values of a single point are stored in a so-called vector. Every vector represents the object that is described by its characteristics. The spatial idea that underlies Vector Space Models does not restrict the objects to tangible items. Indeed, in Distributional Semantics, word meanings are objects, and the characteristics are contexts in which these words appear (Ruette et al. 2014, 212).

When VSMs are combined with the distributional hypothesis, the quantifiable characteristics of the object (i.e. of the word meaning) are the contexts of the word under scrutiny. Parallel to this proposition, VSMs can now also be combined with the *translational equivalence* hypothesis: the quantifiable characteristics which make up an *n*-dimensional space are then the translations or the source language lexemes of a word under scrutiny provided that a relationship of *translational equivalence* has been established (which will be done via the SMM++) between the translation/source language word and the word under study.

The attraction of the VSMs for semantic research resides in the fact that they can be used to quantify semantic similarity "by applying the spatial idea that underlies the Semantic Vector Space Models" (Ruette et al. 2014, 213). This works as follows:

> If two objects are very close to each other in an *n*-dimensional Semantic Vector Space, then they are bound to have very similar values on a number of dimensions. If two objects behave alike for a large number of characteristics, represented by the dimensions, they must be very similar to each other, with respect to these dimensions. Given that we assume that the dimensions in Semantic Vector Spaces represent the Distributional Semantics of a lemma, spatial closeness of two words translates into semantic similarity between these words" (Ruette et al. 2014, 213).

Again, the idea that Semantic Vector Spaces can be combined with the distributional hypothesis can be transposed to the translational hypothesis: if we want to know how semantically similar two words are in translated and non-translated language, we can equally measure the spatial proximity between those two words in both varieties. For instance, we can measure the semantic similarity between *stoel* [chair] and *bank* [bench] in translated Dutch and compare this relationship to the semantic similarity between those same two lexemes in non-translated Dutch. In translated Dutch, *stoel* [chair] and *bank* [bench] are translations and each lexeme is represented by a vector containing all possible source language words obtained from a corpus (as frequency values). For non-translated Dutch, *stoel* [chair] and *bank* [bench] are source language lexemes and each lexeme is represented by a vector containing all possible translations obtained from a corpus (as frequency values). Following the idea that "spatial closeness of two words translates into semantic similarity between these words" (Ruette et al. 2014, 213), we can compare the distances between *stoel* [chair] and *bank* [bench] in both varieties and consequently compare the semantic similarity between the two lexemes for both translated and non-translated Dutch.

In a large, corpus-based study such as ours, each translation or source language lexeme will be represented as a row in a frequency table and each characteristic of the *n*-dimensional space (source language lexeme or translation) will be represented as a column variable in a data matrix. If one wants to see "what kind of semantics" (Heylen et al. 2012, 17) is hidden within such potentially huge data matrices "an intuitive interface to explore the semantic structure captured by an SVS" (Ibid.) will be needed. Such an interface (a visualization) can then be obtained via statistical analysis of those data matrices. In this study, we will apply Correspondence Analysis and Hierarchical Cluster Analysis to yield such visualizations (see section 3.6).

2.4.3 Corpus-based cognitive semantics

In linguistic semantics, thorough quantitative corpus analyses have been combined with theoretical concepts of cognitive linguistics. Heylen et al. compare the work developed by two groups of researchers who have "relatively independently" developed "the methodology of "cognitive linguistically inspired" quantitative corpus analysis" (Heylen et al. 2008, 92): Gries, Stefanowitsch and colleagues on the one hand and Geeraerts, Speelman, Grondelaers and colleagues on the other hand[21]. By integrating cognitive theory into corpus-based approaches to linguistics, the researchers hope to arrive at a more empirical account of lexical meaning. Gries explains that, by bridging the gap between cognitive and corpus-based studies, rather than focusing on the distributional characteristics of different word senses, it should become possible to be informed about "how different word senses are related" (Gries 2006b, 57). The integration of a cognitive linguistic framework within a corpus linguistic study is believed to lead to more "theoretical sophistication" (Gilquin 2010, 16). In this section, we will see that a "prototype model of category structure" is best suited to provide us with the needed theoretical sophistication for this study. This model will make up the theoretical foundation for the interpretations of the obtained visualizations (see chapter 4). The "prototype model of category structure" is considered as one of the important contributions that cognitive semantics has made to the study of word meaning (Geeraerts 2013, 577). In the first part of this section (2.4.3.1), we will zoom in on the notion of prototypicality so that we can use it in an unproblematic way to further describe and interpret our results in the subsequent chapters of this study. In the second part (section 2.4.3.2), we will show how Divjak's proposal to opt for a prototype-based categorization for low-contrastive verbs expressing abstract concepts also seems to be the better choice for our study. In addition, we will comment on her two proposals of internal category organization (schematic or radial structure). Just as Divjak, we will also prefer a radial category organization.

2.4.3.1 A prototype-based view and prototype effects

The development of prototype theory received its most important impetus from psycholinguistic research conducted by Eleanor Rosch and colleagues in the 1970s (Rosch 1975, Rosch & Mervis 1975, Rosch 1978). One of Rosch's most important findings

[21] The comparison between the two approaches will not further be discussed here, but see: Heylen et al. (2008). Briefly, the differences between the approaches situate themselves on the level of the phenomena under investigation, explanatory approaches and the exact statistical technique employed (Heylen et al. 2008, 92-93).

was that “[m]ost, if not all, categories do not have clear cut boundaries” (Rosch 1999 [1978], 196). The idea of fuzzy category boundaries, seemed, however, not easy to connect to the ‘dictate’ of cognitive economy that saw categories as “being as separate from each other and as clear-cut as possible” (Rosch 1999 [1978], 196). Rather than intending to achieve cognitive economy via “formal, necessary and sufficient criteria for category membership”, one could, alternatively, opt to marry fuzzy boundaries with cognitive economy by “conceiving of each category in terms of its clear cases rather than its boundaries” (Rosch 1999 [1978], 196). Prototypes of categories are then “the clearest cases of pry membership defined operationally by people’s judgments of goodness of membership in the category” (Ibid.). Rosch thus considered perception of typicality difference and hence also degree of prototypicality as an empirically verified fact. Given this empirical fact, Rosch went on to ask precisely “what principles determine which items will be judged the more prototypical?” (Rosch 1999 [1978], 197). Her hypothesis was that “prototypes develop trough the same principles such as maximization of cue validity and maximization of category resemblance as those principles governing the formation of categories themselves” (Ibid.). Support for this hypothesis can be found in Rosch & Mervis (1975) who showed that “the more prototypical of a category a member is rated, the more attributes it has in common with other members of the category and the fewer attributes in common with members of the contrasting categories” (Rosch 1999 [1978], 197).

Outside the field of psycholinguistic research, Rosch’s findings have further evolved and influenced psycholexicology on the one hand, and from the mid-1980s (general) onwards also linguistics (Geeraerts 2013, 578). As far as cognitive linguistics is concerned, prototype theory is even seen as one of its cornerstones (Geeraerts 2006 [1989], 145). According to Geeraerts, within linguistics, Rosch’s conclusions that “perceptually based categories do not have sharply delimited borderlines” developed into “a more general prototypical view of natural language categories, more particularly, categories naming natural objects” (Geeraerts 2013, 578). Geeraerts further summarizes the application of prototype theory to the domain of linguistics as follows:

> The theory implies that the range of application of such categories is concentrated round focal points represented by prototypical members of the category. The attributes of these focal members are the structurally most salient properties of the concept in question; conversely, a particular member of the category occupies a focal position because it exhibits the most salient features (Geeraerts 2013, 578).

The importance of the introduction of the notion of prototypicality in linguistic theory lies in the fact that categories do not ‘need’ to be described any longer by lists for necessary and sufficient properties, but can instead be described according to more central and more marginal category members (Gilquin 2006, 160-161). Prototypicality was furthermore extended beyond concrete objects to more abstract categories such as past tense and syntactic constructions (Taylor 1989).

The use of the notion within linguistic theory is, however, not uncontroversial. Geeraerts shows that prototypicality is itself "a prototypical notion with fuzzy boundaries" (Geeraerts 2006 [1989]). Prototypicality, according to Gilquin, then needs to be considered as a "multi-faceted concept":

> bringing together (1) theoretical constructs from cognitive literature and relying on deeply-rooted neurological principles such as the primacy of the concrete over the abstract, (2) frequently occurring patterns of (authentic) linguistic usage, as evidenced in corpus-data, (3) first-come-to mind manifestations of abstract thought, as revealed through elicitation tests and (4) possibly other aspects that contribute to the cognitive salience of a prototype (Gilquin 2006, 180).

By defining prototypicality along these different lines, Gilquin tries to incorporate the four hypotheses uttered by Geeraerts (2006 [1988]) as possible answers to the question: "where does prototypicality come from?". These four hypotheses run as follows: First, the physiological hypothesis: prototypicality is considered as the result of the physiological structure of the perceptual apparatus (Rosch 1973). The problem with this hypothesis is that it is difficult to apply to concepts without physiological basis (Geeraerts 2006 [1988], 28). Second, the referential hypothesis: prototypicality as the result of the fact that "some instances of a category share more attributes with other instances of the category than certain peripheral members of the category" (Geeraerts 2006 [1988], 28). This hypothesis is also referred to as the "family resemblance model of prototypicality" (Rosch & Mervis 1975). The number of shared attributes among the objects, events,... a concept can refer to allow the researcher to compute differences in salience (Geeraerts 2006 [1988], 29). Thirdly, according to the statistical hypothesis, the prototype is that member of a category which is most frequently experienced. Geeraerts (2006 [1988], 29) adds that the second and the third hypothesis can be combined: one can ascribe weights to category attributes on the combined basis of family resemblance and relative frequency (Rosch 1975). Finally, the fourth hypothesis is the psychological (also called functional) hypothesis which states that "it is cognitively advantageous to maximize the conceptual richness of each category through the incorporation of closely related nuances into a single concept because this makes the conceptual system more economic" (Geeraerts 2006 [1988], 29).

We follow Gilquin in her "multi-faceted" view on prototypicality, which incorporates Geeraerts four hypotheses. We will nevertheless be confronted with the following question: if we take a prototype-based view on language in this study – while keeping in mind the possible multiple sources of prototypicality – and we want to make claims about the semantic relationships within the semantic fields presuming a prototype-based organization, how can we be sure that our chosen method will actually render a prototype-based structure? Given the corpus-oriented scope of this research, the most straightforward way of 'ensuring' that the yielded semantic fields will be prototype-based is to integrate both the second (family resemblance / salience) and the third (statistics) hypothesis. In this way, we unite a cognitivist view on prototypicality – "cognitivists tend to consider the prototype as the cognitively most salient exemplar" (Gilquin 2006, 159) – with a corpus-linguistic view which usually considers the prototype as the most frequently corpus-attested item (Gilquin, Ibid.). As Gilquin points out, most of the time, both cognitivists and corpus-linguists assume that salience and frequency coincide with one another (Gilquin, Ibid.). Although Gilquin does not negate the role of frequency in prototypicality, she also cites Sinclair (1991, 36) who argues that "for common words, as a rule, the most frequent meaning is not the one that first comes to mind". In our study, we will then not only take frequency as a measure of prototypicality, we will also propose a way to operationalize salience, and we will do so by taking into account the number of overlapping translations. By doing so, we also tackle the problem that "[t]he lack of convergence between salience and text frequency [could] challenge[ ] the ability of corpora to serve as a shortcut to cognition" (Arppe et al. 2010, 9). By considering translations as attributes, we can apply Geeraerts idea (2006 [1988], 29) that the number of shared attributes (overlapping translations) can be used to compute salience. The principle of *overlap* will be further developed in section 3.3.3. In short, we combine the use of frequency – the statistical hypothesis – and overlap – our operationalization of salience – to determine the status (more prototypical or more peripheral) of the member(s) of the semantic field we plan to visualize.

Geeraerts' four hypotheses can be linked to a number of prototype effects. Just as Rosch was interested in the principles governing prototypicality judgment, within linguistics too researchers felt the need to differentiate between different phenomena that were all linked in some way to prototypicality (or to one of the previously cited hypotheses about the origins of prototypicality) and consequently prefer to talk about prototype effects rather than about prototype theory (Geeraerts 2013, 578). Geeraerts sums up a list of four characteristics about which there exists a consensus in the literature on the fact that "these characteristics are prototypicality effects [...] may be exhibited in various combinations by individual lexical items, and [...] may have very different sources" (Geeraerts 2013, 578). The list of prototypicality effects is determined as follows by Geeraerts:

> First, prototypical categories exhibit degrees of typicality: not every member is equally representative for a category. Second, prototypical categories exhibit a family resemblance structure, or more generally, their semantic structure takes the form of a radial set of clustered and overlapping readings. Third, prototypical categories are blurred at the edges. Fourth, prototypical categories cannot be defined by means of a single set of criterial (necessary and sufficient) attributes (Geeraerts 2010, 187).

The existence of these prototypicality effects will need to be taken into account in the development of our methodology (see chapter 3). If we believe that not every member is equally representative for a category, our method will need to be able to inform us about member representativity (this will be done by calculating the distance from each lexeme to its cluster's centroid, see section 3.6.3). As far as the family resemblance structure is concerned, we have explained in this section how we plan to integrate it in our method, i.e. by means of the so-called *overlap* principle. We will also have to formulate an answer as to how we will deal with the fuzziness of category boundaries (we will impose a minimum threshold for the overlap criterion to somewhat limit the fuzziness (see section 3.4.3) and we will evaluate the remaining fuzziness by assessing the distance of each lexeme to its cluster's centroid as well as to the *centroids* of other clusters (see section 3.6.3)). Lastly, the lexeme selection technique based on the SMM takes translations as its attributes – so categories do not need to be defined according to their necessary and sufficient attributes.

2.4.3.2 A prototype-based categorization of verbs

Divjak remarks that many of the experiments about prototype categorization have been conducted on nouns, so that "[e]xtending prototype categorization to verbs [...] presupposes that knowledge about structures pertaining to nouns might be operative in verbs" (Divjak 2010, 150). Given a number of differences between nouns and verbs – verbs are not stable/ time independent, verbs name intangible events, verbs render relational concepts (Ibid.) – it is indeed plausible that "conceptual categories associated with verbs and adjectives function differently from those associated with nouns (Ibid.). According to Divjak, verbs are in general more abstract concepts than nouns and therefore less tangible, making it more difficult to capture them in prototype representations (Ibid.). As far as the intangibility of the verb concepts is concerned, Divjak (2010, 152) refers to Pulman (1983, 114) who states that verbs will require "more complex and more abstract attributes" than more tangible concepts expressed by nouns (where the prototypical members are those which share most attributes with some members of a category and only some attributes with other, peripheral members). Despite these differences, Divjak indicates that there is "some psychological evidence that people categorize event-related and object-related information in a similar way" (Divjak 2010, 151). There seems to be no doubt however that "categories for intangible relational concepts also display prototype effects" (Ibid., 153), as is shown by Schmid (1993), Taylor (1995, 2003) and Geeraerts (1985, 1988, 1990) (Divjak 2010, 153). Divjak concludes that choosing categorization by prototype is "quite adequate for modeling low-contrastive verbs, expressing abstract concepts such as intention, attempt or result [...]" (Ibid., 150).

Since the semantic domain we aim to investigate in this study also expresses a rather abstract concept (inchoativity), we believe that the above line of reasoning in favor of prototype-based categorization also holds for our study. Divjak herself uses ID tags to set up behavioral profiles for each of the verbs in her study for prototype identification (Divjak 2010, 158). Our own proposition to operationalize translations as attributes might offer an alternative solution to the 'problem' of the complexity of (abstract) verb attributes: an identical type of attributes can be assigned to nouns, verbs and adjectives alike (i.e. their corresponding translations, see chapter 3).

Once we have accepted that a prototype-based organization for the internal structure of a category is a defendable choice, the second question that comes to mind is: what does it look like? (Divjak 2010, 149). According to Divjak, "[w]ithin cognitive linguistics, complex categories are typically represented in one of two ways, i.e. as having a schematic or a radial structure" (Divjak 2010, 149). The first way of representing complex categories follows Langacker's idea of a "schematic network of interrelated senses" (Langacker 1987, 369, 371), where a schema is "an abstract characterization that is fully compatible with all the members of the category it defines" (Divjak 2010, 149). The second way of representing complex categories is as a radial structure (Lakoff 1987, 84): "[a] radial structure is one where there is a central case and conventionalized variations on it which cannot be predicted by general rules". Although both types of categorization "are inherently related aspects of one and the same phenomenon and are often difficult to distinguish in practice" (Langacker 1987, 371 ff; quoted by Divjak 2010, 149), they are different in the sense that schematic networks require full compatibility with all the category members (a checklist of necessary and sufficient attributes), whereas radial category structures are prototype-based, implying that there are degrees of membership (Divjak 2010, 150). Because of the compatibility of the radial category structure with the idea of a prototype-based organization of the internal structure, we will also aim to represent our visualizations as radial structures.

2.5 Conclusion

In this chapter, we have seen that empirical studies of meaning are rather scarce in CBTS. Within the *translation universals* paradigm, for example, the question whether *universals* exist on the semantic level too has not often been raised. We attributed this lack of empirical studies of meaning to the typical status of meaning in translation, i.e. meaning as the invariant of translation. We showed that this alleged invariance of meaning cannot be accepted as a given and that investigating meaning in translation could potentially help us answer the perennial question of the difference between translated and non-translated language. Universal tendencies such as *levelling-out* and *normalization-shining through* are in fact well suited to investigate meaning relationships in translation and such studies could indeed even inform the *universals* research on an explanatory level.

We have put forward the *semantic mirrors method*, which uses translational corpora and integrates *back-translation* to arrive at a selection of lexemes pertaining to the same semantic field. The technique has the potential to lay bare meaning relationships while taking into account the distinction between translated and non-translated language we are so much interested in.

With the prospect of elaborating a bottom-up statistical visualization method for semantic fields in translated and non-translated language, a number of theoretical notions from corpus-based semantics were further explored.

Our envisaged method will contain the following elements: it will apply (a version of) the SMM, it will rely on a WSD-based interpretation of the notion of *translational equivalence* (making the concept operationalizable in a way that is acceptable for research in TS), it will employ statistical visualization techniques that are usually employed in distributional semantics, it will take a prototype-based view on meaning to interpret the statistical visualizations we aim to create.

In order to apply the SMM for our study, however, two practical issues still need to be solved. First, we need to find a way in which the SMM can be applied to retrieve comparable sets of translated/target language on the one hand and sets of original/source language on the other hand (a clear distinction between those sets is of paramount importance, while comparability stays a prerequisite). A second point of attention which cannot be solved by merely applying the SMM is the objective visualization of the results: how to practically create the statistical visualizations of those retrieved sets of lexemes? These two issues will be at the center of the methodology described in the next chapter.

Chapter 3
Methodology

3.1 Introduction

Our first concern in this methodological chapter is to find a way (a technique) to visualize semantic fields in translated and non-translated language. In the previous chapter, we have introduced the SMM, a technique that was originally designed by Dyvik to derive large-scale semantically classified vocabularies for machine translation and other kinds of multilingual processing. We concluded that this technique could potentially offer us a methodological solution for meaning investigation in translation. In this chapter, we want to further explore the SMM and see how the technique can now be employed to compare semantic relationships in translated and non-translated language. We will propose two extensions to the SMM so that the technique can be used to both select (via bottom-up retrieval) and statistically visualize (by measuring the meaning relationships between the lexemes in terms of distances) sets of lexemes as representations of semantic fields of translated language and non-translated language. These visualizations then need to enable us to compare the created semantic fields to each other in such a way that answers to our research questions with respect to the universals of *normalization-shining through* and *levelling out* can be formulated.

In section 3.2 we will present the corpus that will be used in this study, the Dutch Parallel Corpus. In section 3.3, we will give a detailed account of the SMM – as it has been developed by Dyvik. In the next section (section 3.4), we will explain our own extensions of the technique. The first extension is concerned with the integration of translation direction and the asymmetry of translation into the retrieval task; the second extension will focus on how the output of the retrieval task can be used as an input for a statistical visualization of a semantic field. In section 3.5, we will apply the first extension of the SMM to retrieve data sets for the semantic field of *beginnen*/inchoativity in Dutch. In section 3.6, we will apply the second extension of the

SMM by exploring a number of statistical methods that will allow for the visualization of semantic fields. In section 3.6.1, we will carry out a first visual exploration of the data on the basis of *correspondence analysis* before we propose, in section 3.6.2, to carry out a *hierarchical agglomerative clustering* upon the output of the CA. This section also covers the choice of the distance measure (section 3.6.2.1), clustering algorithm (section 3.6.2.2) and number of clusters (section 3.6.2.3) for the HAC. In the final part of this section (section 3.6.2.4), we will compare the chosen procedure (CA on a HAC, Euclidean distance, Ward's Minimum Variance Method) to alternative combinations of distance measures, clustering algorithms and spatial maps by assessing the overall strength of the cluster structures of those combinations.

In section 3.6.3. we present a methodological solution – the calculation of the distances of the clusters to the *centroid* of the semantic space and calculation of *medoids* – to investigate the (changing) prototype-based organization of meaning distinctions within semantic fields of translated and non-translated Dutch (semasiological *levelling out*) and of lexemes within the meaning distinctions (onomasiological *levelling out*) revealed by the clusters.

In section 3.6.4, methodological solutions are proposed to investigate *shining through* on both the semasiological and the onomasiological level.

3.2 The Dutch Parallel Corpus

Rather than compiling our own corpus, we decide to work with an existing corpus; the Dutch Parallel Corpus (DPC). The DPC was developed as part of the STEVIN[22] program. The primary goal of this program was "to set up an effective digital language infrastructure for Dutch, and to carry out strategic research in the field of language and speech technology for Dutch" (Spyns 2013, 1). The DPC is a ten-million-word, sentence aligned, both parallel and comparable corpus (it is *de facto* a parallel corpus which can also be used as a comparable corpus). Within Laviosa's terminological apparatus (presented in section 2.2.1.3), the DPC can be described as a multi-source, parallel multilingual corpus. 'Multi-source' since Dutch, French and English can all three be the source language of the texts in the corpus (and also the target language); 'parallel' because the texts in one language are the originals of the translations in the other language; and 'multilingual' because more than two languages are involved.

[22] STEVIN is the Dutch acronym for "Essential Speech and Language Technology Resources for Dutch".

The DPC offers a number of indisputable advantages. With respect to corpus size the DPC is, to our knowledge and at the time of writing, the largest available parallel corpus of Dutch. It is furthermore 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). Only for the text type 'literary texts', the corpus is not strictly balanced according to translation direction, but 'only' according to language pair (Paulussen et al. 2013, 187). The five text types on the 'superordinate level' are further subdivided into 19 'basic levels', but the latter have "no further implications for the balancing of the corpus" (Macken et al. 2011, 378). Each text type accounts for 2,000,000 words and within each text type, each translation direction contains 500,000 words (Macken et al. 2011, 376-378). All text files consist of written text material (no data carriers other than text files are included), but no distinction is made in the DPC between 'spoken' text material and 'written' text material (Delaere 2015, 59), although available meta-data indeed allow the user to identify the 'spoken' text material as such and to distinguish between texts "written to be read", "written to be spoken" or "written reproduction[s] of spoken language" (Ibid.). It is important to keep in mind that the 'spoken' text material in the DPC is categorized under the superordinate text type level 'administrative texts' (Ibid.), together with 'written' text material. Divergent results for the text type 'administrative texts' in a corpus study focusing on genre specific phenomena could thus be due to the invisible inclusion of this parameter into the text type. The DPC further offers the possibility to differentiate between "regional language varieties" (Delaere 2015, 48) such as Belgian Dutch and Netherlandic Dutch, Belgian French and French French and British English and American English. It is also important to add that the DPC is built up of complete texts, not of samples and that we are dealing here with a 'closed' corpus, meaning that no data are added any further to the corpus.

The DPC indeed fulfills all the prerequisites to be a 'representative corpus' with regard to corpus size, content and types of text files (see section 2.2.1). The corpus is aligned on the sentence level (the alignment was carried out by a combination of three alignment tools) (see Paulussen et al. 2013, 190-191 for more details on the different tools, their advantages and drawbacks). The DPC is furthermore enriched with linguistic annotations such as part-of-speech tagging and lemmatization (Paulussen et al. 2013, 191). With regard to lemmatization, Macken et al. (2011, 384) mention an average accuracy rate for lemmatization of 97.6%. Delaere (2015, 50) remarks that for the Dutch data (displaying an average lemmatization rate of 96,5%), this implies that "for each 1.7 sentences, 1 word is lemmatized erroneously" (Delaere 2015, 50). We agree with Delaere that it is important to keep in mind that "these results may have influenced the output results of our corpus queries" (Delaere, Ibid.), since our queries rely on lemmas. On the other hand, it should be noted that an average accuracy score of 97.6% is considered (more than) acceptable; part-of-speech taggers, for instance, usually reach accuracy rates around 95% (Macken et al. 2011, 383), so any scholar who uses part-of-speech tagged and/or lemmatized corpora will be faced with the same 'handicap' of imperfect lemmatization.

The official web-interface of the DPC[23] displays the results of a search query as concordanced observations. The research group within which this study has been carried out developed an own interface. For this study, we used the very user friendly "graphical search engine" developed as part of the COMURE project to access the DPC[24]. The search engine offers the following search options: language (one can select one or several sub-corpora of regional language varieties), word form (one can search one specific word, or a combination of words; searches can also be carried out via regular expressions), lemma (by querying the lemmatized form, one obtains all word forms of the lemma), part-of-speech (the search can be based on or reduced by the morphosyntactic class of a word), attributes (additional information obtained by the part-of-speech tagging can also be queried) and frequency (the frequency with which a queried word, lemma or part-of-speech occurs in a sentence can be determined, including the possibility of negative searches) (Delaere 2015, 62-65).

Finally, Delaere's thorough investigation of the DPC laid bare a number of problem areas which were not pointed out by Paulussen et al. (2013) or Macken et al. (2011). Especially the 'basic-levels' of the sub-corpora seemed problematic: the labeling on this level appeared rather often erroneous or absent, and little information was given with regard to the selection of the texts pertaining to each of the basic levels (Delaere 2015, 52). It can also be added that the term 'basic level' is prone to confusion with the prototype-theoretical term 'basic level categories'. In addition, Delaere reported that for about 9% of the texts, the source language appeared to be unknown. While the first problem of 'basic-level' annotation is of little importance to this study, the second issue is indeed more problematic since source language and target language need to be selected at each step of the SMM++. Given the extreme difficulty of retrieving the source language of a given text post hoc, we decided to discard those observations for which the DPC does not indicate the source language (source language 'unknown').

[23] Access to the demo version via http://dpc.inl.nl/indexd.php

[24] Access to the full version (password required) via http://dpcserv.ugent.be/comure/

3.3 The Semantic Mirrors Method

In the previous chapter, we have introduced the notion of Semantic Mirroring. We concluded that the SMM, which was originally designed to derive large-scale semantically classified vocabularies, has the potential to lay bare meaning relationships in translated and non-translated language. We have explicated – on a theoretical level – how the technique works and we have illustrated its usefulness with some examples from contrastive studies. Crucially, the technique of Semantic Mirroring is based on the following assumption:

> [S]emantically 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).

In this section, we will first present the 'work flow' of the SMM as it was proposed by Dyvik (3.3.1). After this description of the different stages of the SMM, we will take a step back and explore the prerequisites and assumptions one needs to take into consideration *before* an SMM can be carried out (3.3.2). We will further explicitate the rationale behind the *Overlap Threshold* (in section 3.3.3) as a crucial element of the technique which ensures that semantically related lexemes can be separated from semantically unrelated ones.

3.3.1 Work flow of the SMM

Dyvik starts from an initial polysemous *lexeme a in Language A* and extracts all its translations in Language B manually from the English-Norwegian Parallel Corpus (ENPC), a sentence-aligned corpus. He calls this set of translations the *first T-image of a in Language B*[25]*.* Then, commensurably, the translations back in 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, we could represent the work flow as follows:

[25] For the sake of clarity, we have added the adjective "first" here. "*The First T-image*" thus refers to what Dyvik himself calls "*the t-image*". The "*Inverse t-image*" and "*Second t-image*" are the exact names given by Dyvik to the following steps in the SMM.

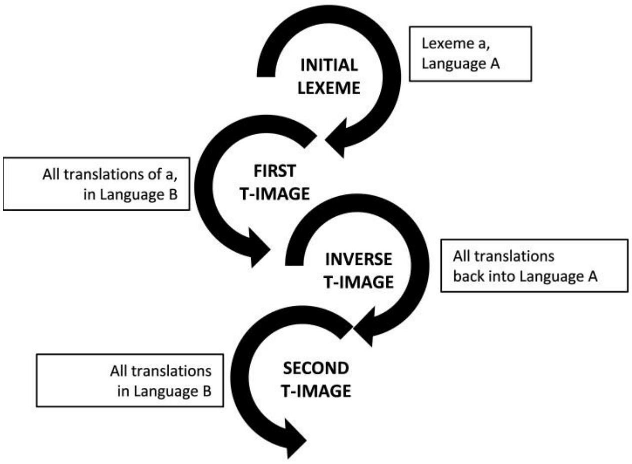

INITIAL LEXEME
Lexeme a, Language A
All translations of a, in Language B
FIRST T-IMAGE
INVERSE T-IMAGE
All translations back into Language A
All translations in Language B
SECOND T-IMAGE

Figure 8 Work flow of the SMM

3.3.2 Prerequisites and assumptions

A practical prerequisite to carry out the technique 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).

From the corpus which has been chosen, the researcher needs to be able to extract a set of alternative translations for each lemma one wishes to investigate (Dyvik 2005, 31). After the application of the different steps of the SMM, this will ultimately create a "network of translational correspondences uniting the vocabularies of the two languages" (Ibid.). Based on Dyvik's ideas, which we amply discussed in the previous chapter, and based on the following assumptions (verbatim from Dyvik (2005, 31-32)), the created network will allow us to use "each language as the 'semantic mirror' of the other". The assumptions Dyvik puts forward are as follows:

(1) Semantically closely related words tend to have strongly overlapping sets of translations.
(2) Words with wide meanings tend to have a higher number of translations than words with narrow meanings.
(3) If a word *a* is a hyponym of a word *b* (such as *tasty* of *good*, for example), then the possible translations of *a* will probably be a subset of the possible translations of *b*.
(4) Contrastive ambiguity, i.e., ambiguity between two unrelated senses of a word, such as the two senses of the English noun *band* ('orchestra' and 'piece of tape'), tends to be a historically accidental and idiosyncratic property of

individual words. Hence we don't expect to find instances of the same contrastive ambiguity replicated by other words in the language or by words in the other languages. (More precisely, we should talk about ambiguous *phonological/graphic* words here, since such ambiguity is normally analysed as homonymy and hence as involving two lemmas.)

(5) Words with unrelated meanings will not share translations into another language, except in cases where the shared (graphic/phonological) word is contrastively ambiguous between two unrelated meanings. By assumption (4) there should then be at most one such shared word (Dyvik 2005, 31-32).

3.3.3 Overlap threshold

The first step that needs to be taken, is to isolate "mutually unrelated senses of each word" (Dyvik 2005, 32). For this, the resulting lexemes of the *first T-image* are used. We will try to illustrate the difference between "related word senses", "unrelated word senses" and "mutually unrelated word senses" with the example of the Dutch word *bank* (Figure 9), which can be translated in French as *institution financière* [financial institution], *banque* [financial institution], *banc* [seat] and *fauteuil* [armchair]. This distinction between the different types of senses is not explicitly made by Dyvik, but can be derived from the procedure he proposes for word sense isolation. Our explanation can furthermore be helpful to come to a better understanding of the possible pitfalls of *back-and-forth* translation and simultaneously reveal how to bypass them. Finally, the distinction will be of importance to our Extension of the SMM.

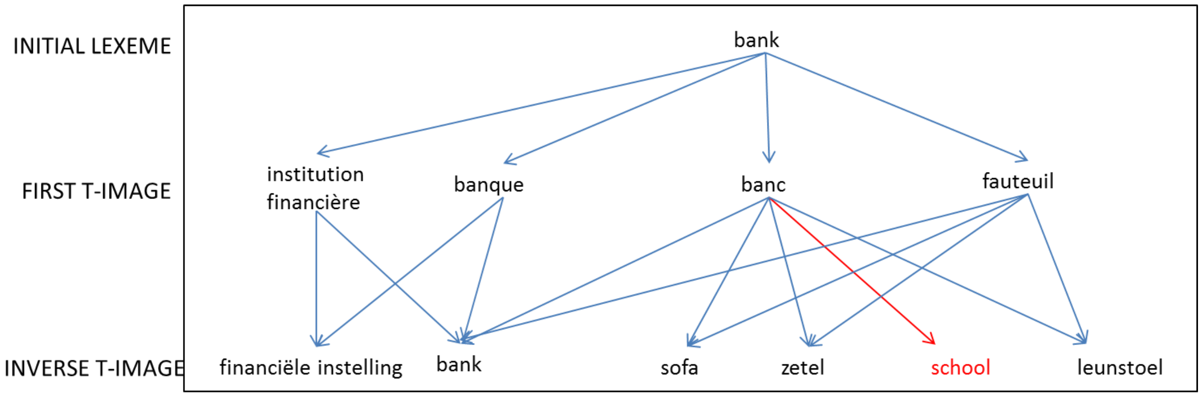


Figure 9 Example of the (ficticious) SMM of *bank*

3.3.3.1 Unrelated word senses

When we take a look at the translations back into Dutch (the *inverse T-image*) of *banque* and *banc*, we see that these lexemes only share the *initial lexeme bank* itself in the *inverse T-image*. *Banque* (Figure 10) is connected in the *inverse T-image* (i.e. 'can be translated back into Dutch as') to *bank* and *financiële instelling.* Additionally, we could also say that the *inverse T-images bank* and *financiële instelling* are semantically related to each other (via *banque):*

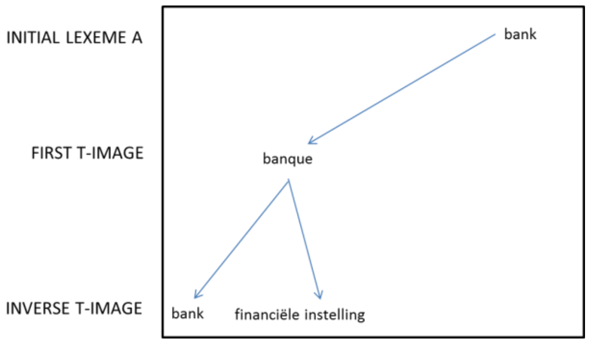


Figure 10 Inverse T-image *of banque*

Banc (Figure 11) on the other hand, is connected in the *inverse T-image* to *bank, sofa, zetel* and *leunstoel,* which means that the *inverse T-image* lexeme *bank* is semantically related to the other *inverse T-image* lexemes *sofa, zetel* and *leunstoel* (via *banc):*

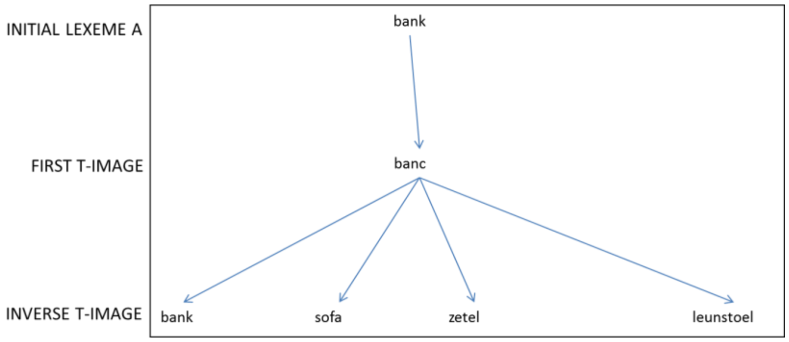


Figure 11 Inverse T-image of *banc*

The *first T-images banque* and *banc* only share *bank* on the level of the *inverse T-image,* so *banque* and *banc* are "not directly connected by means of intersections with other sets" (Dyvik 2005, 32) indicating that their semantic relatedness cannot be proven (and that *bank* is contrastively ambiguous between French *banque* and *banc*). This observation corresponds with Dyvik's assumption (4): the Dutch lexeme *bank* is indeed *homonymous* between *bank* "financial institution" and *bank* "seat". We also see evidence here for Dyvik's assumption (5): the words *banque* and *banc* indeed only share ("at most") one word (translation) at the level of the *inverse T-image*, i.e. the contrastively ambiguous *bank.* We can conclude that an *initial lexeme* (e.g. *bank*) possesses two distinct, unrelated senses (e.g. "financial institution" and "seat") if the only shared word between their two sets of lexemes in the *inverse T-image* is the *initial lexeme* (which is the case here: the two sets only share *bank*).

3.3.3.2 Related word senses

When we now look at the *first T-images banc* and *fauteuil* (Figures 12 and 13), we see that *banc* is connected to *bank, sofa, zetel* and *leunstoel* in the *inverse T-image* (Figure 12), and that *fauteuil* is connected to *bank, sofa, zetel* and *leunstoel* in the *inverse T-image* (Figure 13). We can state that, in their *inverse T-images*, *banc* and *fauteuil* share, apart from *bank,* also *sofa*, *zetel* and *leunstoel*. *Banc* and *fauteuil* are thus directly connected by means of intersections with other sets: they do not only share *bank* in the *inverse T-image*, they also share *sofa*, *zetel* and *leunstoel*, proving the closer semantic relatedness of *banc* and *fauteuil*, and also showing that *bank*, *sofa*, *zetel* and *leunstoel* are semantically related.

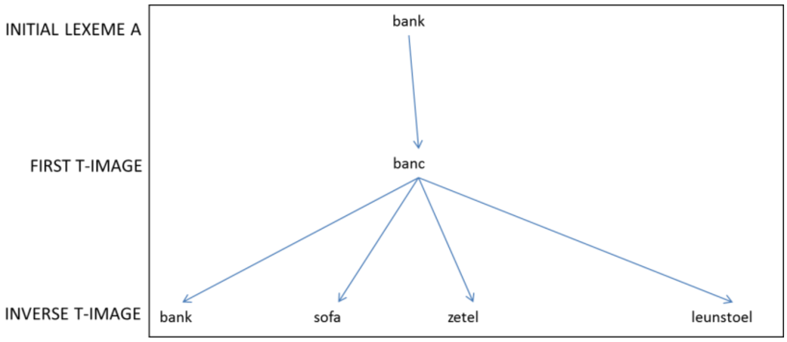


Figure 12 Inverse T-image of *banc*

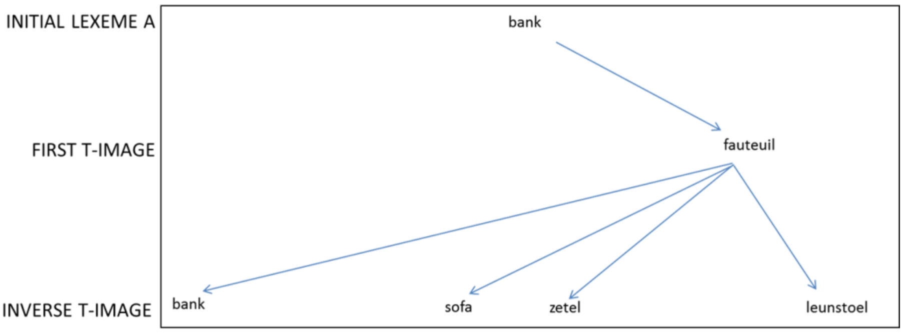


Figure 13 Inverse T-image of *fauteuil*

3.3.3.3 Mutually unrelated word senses

A final possible scenario concerns the example of the Dutch word *school* [school] in the *inverse T-image* (look back at Figure 9, the example of the (fictitious) SMM of *bank*). *School* is a possible translation of the French *inverse T-image* word *banc* back into Dutch, in its meaning “school of fishes”. But this latter meaning is not a meaning of *bank*. Without any knowledge of Dutch and French, the unrelatedness can be deduced from the *translational relation: school* is only translationally related to its French source lexeme *banc*, but it is not *related* to *bank* on the level of the *inverse T-image,* implying that the senses of *bank* and *school* are mutually unrelated (Figure 14). Whereas *unrelated senses* shared only their initial lexeme in the *inverse T-image* – enabling us to distinguish two distinct, unrelated senses of the initial lexeme *bank*; *mutually unrelated senses* such as *school* and *bank* are not at all related to each other in the *inverse T-image*.

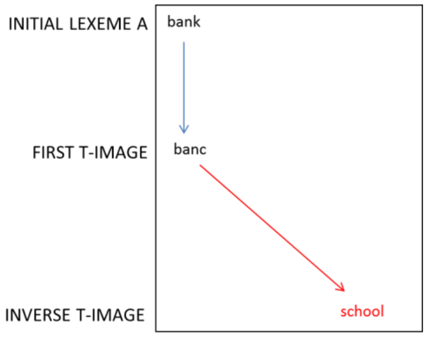


Figure 14 Mutually unrelated sense *school*

3.3.3.4 Word sense individuation

The individuation of word senses can now take place: one of the meanings of *bank* can be expressed by *bank* and *financiële instelling,* another meaning of *bank* can be expressed by *bank, sofa, zetel, leunstoel.* These two meanings are unrelated to each other, they form different semantic fields. *School* is not a sense of the initial lexeme *bank* and should be disregarded for the further investigation of the senses of *bank.* Dyvik summarizes the principle on which the isolation of word senses takes place as follows:

> In our translational approach, the semantic fields are isolated on the basis of *overlapping t-images [first T-images]*: two senses belong to the same semantic field if they have intersecting first *t*-images (after sense individuation one member in the intersection is sufficient), or if there is a sequence of such intersecting *t*-images *[first T-images]* joining them (Dyvik 2005, 33, our emphasis, our own terminology is added between brackets for clarity's sake).

If one is interested in studying one specific semantic field, a criterion of *overlapping (first) t-images* or *overlap* can be observed, meaning that a lexeme at the level of the *inverse T-image* is only selected when it is related to at least two lexemes on the level of the *first T-image.* In this way, for the example of *bank,* we see that *school* is linked to only one lexeme on the level of the *first T-image* viz. *banc*. *School* does not meet the *overlap criterion,* which is an indication that it pertains to a different semantic field. As for *sofa,* for example, we see that it is linked to both *banc* and *fauteuil* on the level of the *first T-image,* proving that it pertains to the semantic field under scrutiny.

By consequence, we could say that, by taking into account a criterion of *overlap* between the *inverse T-image* lexemes and the *first T-image* lexemes (every lexeme selected on the level of the *inverse T-image* must be a translation of at least two *first T-image* lexemes), it is guaranteed that *mutually unrelated senses* are excluded. If words without *overlap* were included in the analysis (i.e. words which are not related to at least two lexemes on the level of the *first T-image*), the result of the SMM would risk to contain senses which are mutually unrelated, meaning that they are in fact not a sense of the word under study.

3.3.3.5 Necessity of overlap

The previous paragraphs have shown that *overlap* is a crucial notion for the selection of those lexemes which pertain to the same semantic field. It has also been shown that the existence of more than one translation for a given word, is not a sufficient argument to accept that the word is ambiguous (Dyvik 2005, 30). In fact, it only implies that the denotation of the word spans the denotations of two words in a different language (p.29). This observation has important implications for the use of the *translational relation* for meaning investigation: "non-transitive translational connections may tie together semantically distant words in the same semantic field" (Ibid.) – as we have shown in the example of *school*. Dyvik makes an important point about the use of *back-translation* in this regard*:* the *translational relation* should be used with care when used for the establishment of semantic relatedness, and *overlap* is a necessary criterion if one wants to 'confine' a semantic field. This problem has also been observed in computational linguistics, where it is generally solved by the addition of another language (Lefever et al. 2013). The appearance of *overlapping translations* was already formulated by Ivir (see section 2.3.2 of this study: "each L_2 correspondent will be related to a number of other L_1 items too, besides the L_1 with which the analysis was initiated") but Ivir did, to our knowledge, never exploit this idea explicitly as a validation of the semantic relatedness between the lexemes of a semantic field. Dyvik's point about the semantic informativity of translations makes his technique directly applicable for lexical semantic research. His reflection about what happens to both ambiguous and unrelated senses when the *translational relation* is used via *back-translation* furthermore offers useful insights into what exactly happens when one utilizes translation for meaning-informative tasks.

3.4 Extended Semantic Mirrors Method: SMM++

As mentioned before, we want to find an adequate way to retrieve lexemes as candidate-members of a semantic field under scrutiny and do so for both non-translated (original/source) language and translated (target) language in order to arrive at comparable visualizations of semantic fields of a same initial lexeme in both translated and non-translated language.

The SMM developed by Dyvik, and some of the additions proposed by contrastive linguists who applied the technique answer our retrieval question: by going back and forth between sources and translations, and by creating new sets of data at every stage of the exercise, a set of candidate-lexemes of a semantic field can be obtained. So far, we have found an expansive, meaning informative technique which can be used for the retrieval of lexemes pertaining to a semantic field.

In order to provide a 'complete' methodological answer, the SMM will still need to undergo a few extensions. The SMM can indeed help us to retrieve candidate-lexemes for a semantic field, but if we want to use Dyvik's technique as a methodological tool to investigate translational phenomena (via a comparison of semantic fields of translated and non-translated language) a number of issues need to be dealt with.

In this section, we will propose two extensions of the SMM. Our first extension is concerned with the integration of translation direction and the asymmetry of translation into the retrieval task (3.4.1); the second extension we will focus on how the output of the retrieval task can be used as an input for a statistical visualization of a semantic field (3.4.2).

3.4.1 Extension 1: Translation direction and asymmetry of translation

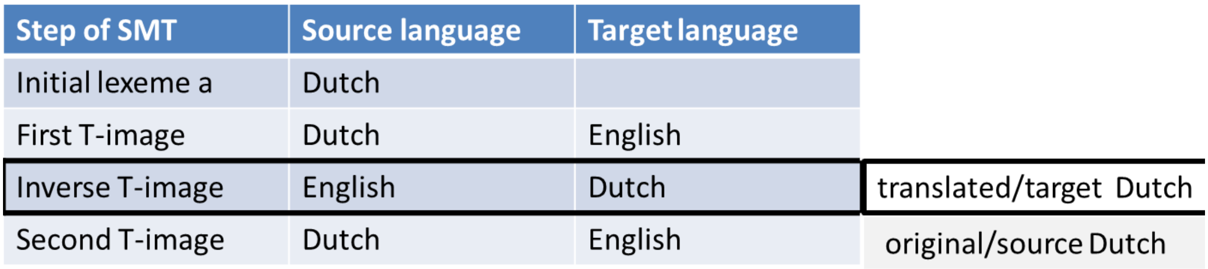
The SMM considers the *translational relation* as symmetric, i.e. a relation which exists irrespective of the translation direction. Recall here that Dyvik's goal is to develop a method for automatic thesaurus building, and, in his capacity of computational lexicographer, he sees and utilizes the *translational relation* as a symmetric relation "disregarding the direction of translation" (Dyvik 2005, 33). The addition of a final step – the *second T-image* – results in a set of Language B lexemes (these are translations into Language B of the Language A lexemes from the *inverse T-image*). Just as the resultant information from the *inverse T-image* (translations into Language A of the Language B lexemes from the *first T-image*) permits the establishment of a semantic field in Language A, the *second T-image* provides the necessary information to establish a semantic field in Language B, and "paired semantic fields in the two languages involved" (Ibid.) are created.

For the translation studies scholar, accepting the symmetry of the translational relation would be refuting almost the totality of the existing research tradition in translation studies. If we want to use the SMM for research in TS, we will inevitably have to take into account the asymmetric nature of the translational relation as well as the reality of translation direction. This implies that, in our view, translation – as an activity which forms the subject of research in TS – always happens in the direction *from* a source language *into* a target language. Differentiating between source and target language does matter in TS, for it is precisely the influence of *either* source or *target* *language* (or both) on the process and the final product of translation which is a pending subject of research in TS.

We therefore want to create two sets of data which can form the basis for a comparison of a semantic field of a lexeme under scrutiny, once in non-translated (original/source) language (in our case non-translated Dutch), and a second time in translated (target) language (in our case translated Dutch with English or French as a source language). The sets representing non-translated Dutch on the one hand and translated Dutch on the other hand need furthermore be (easily) comparable.

To meet this goal, we will look back at the original structure of the SMM as it was conceived by Dyvik, but we will now focus on *translation direction,* keeping in mind that the semantic fields we want to create need to consist of lexemes in the same language as the *initial lexeme,* and that we want to differentiate between non-translated (original/source) language and translated (target) language. Suppose an SMM is carried out on an *initial lexeme* a *in language* A, for which language A is Dutch and language B is English, then the following scheme applies:

Table 1 Source and target language in the different steps of 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

From this Table 1, it becomes clear that Dutch (Language A) is a source language in the *first* and the *second T-image* and a target language in the *inverse T-image*. 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) or disregarding it (like Aijmer and Simon-Vandenbergen 2004), 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 TS research. Whenever we want to investigate *translated language,* we will refer to it as TransLanguage$_{A}$ (in this study TransDutch$_{ENG}$ and TransDutch$_{FR}$); this implies that we are talking about the set of data obtained in the *inverse T-image* with a Language B (in our study English or French) as a source language and any Language A (in our study Dutch) as a target language. Whenever we want to examine *non-translated (original/source) language,* we will talk about SourceLanguage$_A$ (in this study SourceDutch); the underlying data set will be the one obtained in the *second T-image* with any language A (in our study Dutch) as a source language and any Language B (here: English or French) as a target language.

3.4.2 Extension 2: Statistical implementability of the data sets

In the previous section, we have dealt with the asymmetric nature of translation and we have determined a way to compile sets of translated and non-translated language by extending the existing SMM. The next step now is to arrive at comparable visualizations of those sets of lexemes. The information which has so far been obtained only gives the researcher 'sets of lexemes', but does not propose any kind of organization of those lexemes which could further inform us about the semantic relatedness between the lexemes.

Within the original SMM, hierarchical patterns are "only based on overlap relations among *t*-images" and are obtained by ranking the lexemes "independently of frequency of occurrence" (Dyvik 1998, 73). 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, and following Dyvik, the centrality of *bank* in a field with *bank, financiële instelling, sofa* and *leunstoel* could 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* (ficticious) 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

Visualization based solely on overlapping t-images is then realized via Venn diagrams, which tend to get rather complex to interpret, as the following example (Figure 15) by Dyvik (2011) shows:

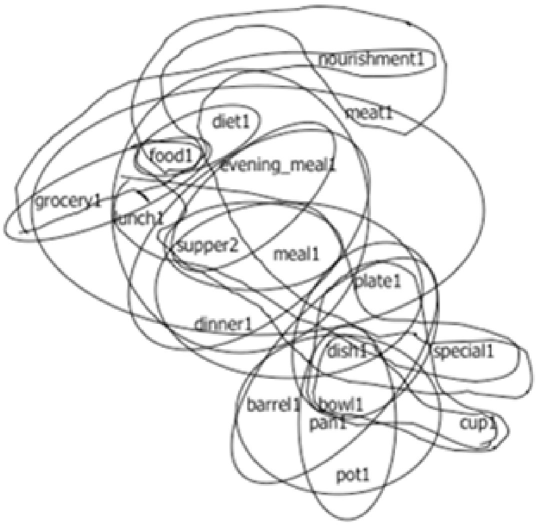


Figure 15 A structured semantic field derived from EPNC, copied from Dyvik (2011)

This apparent 'weak point' of the SMM has led computational linguists to propose different methods of visualization which can be of use for computational research purposes (see e.g. Priss & Old 2005). It is not in the direct scope of our investigation to computationally implement the SMM; instead, we aim to arrive at objective visualizations which can provide us with more insights into the alleged semantic differences between translated and non-translated language. As a result, our goal – and the methods we consequently want to draw on – are more closely connected to distributional semantics. Within that framework, the typical approach is to collect occurrence counts of words and other words/features in a frequency table. The reason is that frequencies indicate the strength of certain relations, i.e. they will tell us which patterns are important. 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 extension. 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	32	45
financiële instelling$_{[nl]}$	121	0	0
sofa$_{[nl]}$	0	98	32
leunstoel$_{[nl]}$	0	67	43

The occurrence counts in the frequency tables implicitly also contain the number of overlapping translations (or source language lexemes). Hence, the frequency tables contain information about both the frequency of co-occurrence of each source language lexeme (or translation) with each translation (or source language lexeme) as well as overlap information about which translations (or source language lexemes) are attested for each source language lexeme (or translation). 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 et al. 1998) and Hierarchical Cluster Analysis (Baayen 2008, 138; Gries 2013, 336), a technique that will permit us to 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 & Gries 2009; Gries 2012; Divjak & Fieller 2014).

3.4.3 Technical finetuning

Although the integration of frequency information into the SMM makes it possible to process the results statistically, one problem still remains. SMM is an expansive technique, implying that, with every step, more and new information is generated, in our case: new translation solutions for the lexeme(s) are retrieved, and their number increases in every step of the mirror analysis. Although this effect is of course at the core of the technique, it also implies that the number of possible translation solutions grows exponentially with every step of the mirror analysis, leading to data sets which are difficult if not impossible (i) to manage manually or even semi-automatically and (ii) to compare with each another (depending on the initial lexeme in Language A one chooses or on the Language B one chooses, the SMM will select different lexemes).

First, let us take a closer look at the problem of how to manage these (ever) expanding data sets within the retrieval task of the SMM. Translators come up with very creative solutions, even in non-fictional, non-literary texts. For example, within a corpus study, this creativity results in the following: for a verb as 'basic' as *beginnen* [to begin], more than 47 different translations in English appear for a total of 382 translational pairs of sentences with *beginnen* in the Dutch source text in the Dutch Parallel Corpus. It can be very interesting, both from a contrastive linguistic as from a translational perspective, to investigate all of these instances, but it would not answer our research question (we are looking for a means to *compare* semantic relationships in translated language and non-translated language). For this reason, we agree with Dyvik to exclude completely *un*predictable translations – translators' idiosyncracies – from our analysis. More specifically, we will apply a frequency threshold of three attestations, allowing us to work with a manageable number of possible translational pairs. Our choice of a frequency threshold of three observations is motivated merely by pragmatic considerations. A first argument to exclude hapax and dis legomena might be found in research by Evert:

> [f]or the time being, however, we must assume that probability estimates and p-values for the lowest-frequency types are distorted in unpredictable ways. [...] these conclusions provide theoretical support for frequency cutoff thresholds. Data with cooccurrence frequency f < 3, i.e. the hapax and dis legomena, should always be excluded from the statistical analysis (Evert 2005, 133).

It must be admitted, however, that 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 in the SMM++ and the more information could be gained from the analysis. With a lower frequency threshold, lexemes which are further away from the 'prototypical center' of the field under study would also have been included in the analysis, an obviously beneficial effect which could ultimately have lead to more insightful visualizations. 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.

A second 'restriction of the data' is necessary if we want to make sure that the sets – of which we have just shown that they can account for translated and non-translated language – are also acceptably comparable. We will therefore respect the following rule of selection for data at the level of the *second T-image* (representing non-translated language): in the *second T-image*, an 'observation' (source-target sentence pair holding the lexeme under investigation) will only be selected when the Language B translation is identical to one of the Language B source language lexemes of the *inverse T-image* (representing translated language). As a result, the row names and column variables of the data matrices in the *inverse T-image* (representing translated language) and the *second T-image* (representing non-translated language) will be identical, their difference will lay in their status. In the frequency table representing non-translated language (SourceDutch), the rows are (Dutch) source language lexemes and the columns are (English or French) translations, in the frequency table representing translated language (TransDutch$_{ENG}$ or TransDutch$_{FR}$), the rows are (Dutch) translations and the columns are (English or French) source language lexemes. Of course, the frequency counts in the tables will also be different (as illustrated by the difference between Table 3 and Table 4 for the fictitious example of *bank*). A similar restriction was also suggested by Dyvik (1998, 60) in order to eliminate those results which are unrelated to the *initial lexeme*. Shortly put: the lexemes which are members of each of the data sets selected for statistical analysis and further visualization are kept identical (the *inverse T-image* provides the lexemes for the semantic field of translated language, and *the second T-image* provides the lexemes for the semantic field of non-translated language), but the 'content' (frequency information and translational status) of the data sets differs since source and target language are in fact inversed in the two sets of data. In this way, we solve the semantic paradox of Krzeszowski (1990) which we are facing here that "what is identical is not subject to comparison, and what is different is not comparable" (Krzeszowski 1990, 7): we propose to select *identical* lexemes, but because of their translational status, both data sets are nonetheless different; solving the paradox and making the two sets of data comparable to each other.

Conclusively, the previously mentioned adjustments will enable us (i) to select a manageable amount of manually controlled data on which a quantitative analysis can be carried out and (ii) to arrive at two data sets which will be justifiably comparable to each other.

3.4.4 Conceptual issue

Following the proposed extensions of the SMM a frequency table is obtained for *non-translated (source)* language on the basis of *translational data* and a *translation-based* method. Admittedly, both the nature of the data as well as the nature of the method could well be held against it. In this section, we show that it can be made conceptually acceptable to use translational data and a translation-based method to obtain a frequency table for non-translated language.

One of the basic assumptions when implementing a method such as the SMM is exactly the idea that the translational relationship can be used as an analytical basis, i.e. we consider "sets of translationally corresponding items across languages as *the primitives of semantic descriptions*" (Dyvik 2005, 31). As a consequence, the translations which are generated by the SMM in the pivot language(s) can be considered as analogous to *semantic features*. These *semantic primitives* or *semantic features* are similar to the *attributes* of the prototype-based theory on semantic organization we presented in chapter 2. If we accept that translations can indeed constitute a kind of attribute, then a semantic description on the basis of translations becomes acceptable and the visualization of non-translated language on the basis of translations (as semantic features) becomes defensible too. The fact that different languages carve up the world in different ways is used to the advantage of the proposed method: contrastive differences can be seen as a reflection of difference(s) (in classification) of semantic properties and can consequently be semantically informative.

Obviously, we cannot escape the fact that, whatever we do to eliminate the effect of translation when investigating non-translated (source) language within our data sets, the data will always remain *yielded by* a translational technique which will inevitably leave some kind of a trace.

As explained in section 3.4.1, we use source language data to investigate non-translated (source) language. Although the *selection* of the data is based on a translation-based technique, viz. the SMM, the observations themselves are source language data on which translation cannot have had an impact (the use of a specific source language lexeme in its non-translated environment cannot have been affected by translation simply because it is not translated). We are aware of the fact that the mere selection of a text as a source text, i.e. a text selected to be translated, already *has* an influence: some texts might be more often and more commonly selected for translation than others, whereas still others may have been excluded due to various factors– these are sometimes referred to as *preliminary norms* (Toury 1995).

One could argue that monolingual data would better fit the purpose of visualizing non-translated language structure. Although this is a valid point, previous studies using monolingual reference corpora have faced major comparability issues due to corpus size or uncertainty about the (translational) status of the texts in the presumed original language corpora (e.g. Förster Hegrenaes 2014). Another option would be to base the intended visualizations on a different hypothesis which does not rely on *translations as semantic features*. If the distributional hypothesis were applied, then only the monolingual contextual information of the Dutch source language sentences would (have to) be used for the visualization of non-translated language[26].

Nevertheless, if we want to make a 'fair' comparison between original language and translated language using the *same* technique, the *same* hypothesis and the *same* data, we can take some additional steps to keep the possible source language influence to a minimum. As a first precautionary measure, we will refer to these data sets and their subsequent visualizations as SourceLanguage$_A$ instead of OriginalLanguage$_A$. Secondly, we will combine the data of two semantic mirrors for the SourceLanguage$_A$ data set. This means that the semantic features from two distinct languages will be combined for the visualization of SourceLanguage$_A$. In this way, we maximize the neutralization of any possible specific influence of the semantic features (translations) on the visualization of SourceLanguage$_A$.

3.5 Applying the first extension of the SMM to retrieve data sets for *beginnen*

In this section, we will apply the SMM++ retrieval task to obtain data sets which can represent the semantic field of *beginnen* / inchoativity in Dutch. The corpus which was used to retrieve the data is the Dutch Parallel Corpus; its structure as well as its data-extraction process are discussed in section 3.2. We will describe the establishment of three resultant data sets by applying the SMM++ to the initial lexeme *beginnen* in the DPC. One data set is obtained for non-translated Dutch (SourceDutch) and two data sets for translated Dutch, one with English as a Language B (TransDutch$_{ENG}$) and a second one with French as a Language B (TransDutch$_{FR}$). All data sets were retrieved following the exact procedure described above. We chose *beginnen* as the initial lexeme, because we consider it as the most prototypical expression of inchoativity: 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 (see Oostdijk et al. 2013).

The first mirroring will be carried out with English as a Language B, the second mirroring with French as a Language B. The *second T-image* of *beginnen* with English as a Language B and the *second T-image* of *beginnen* with French as a Language B will be joined into one the data set SourceDutch. The *inverse T-image* of *beginnen* with English as a

[26] Vandevoorde et al. (2016) show that semantic fields of *beginnen* / inchoativity obtained via the distributional method are similar to those obtained via the translational method.

Language B will result in the data set TransDutch$_{ENG}$, the *inverse T-image* of *beginnen* with French as a Language B will result in the data set TransDutch$_{FR}$.

3.5.1 SMM++ of beginnen$_{ENG}$[27]

3.5.1.1 First T-image of beginnen$_{ENG}$

The attestations of the Dutch verb *beginnen* were queried in the DPC via the interface developed by Delaere (2015, 62). A lemma-based query was carried out rendering all sentences with *beginnen* in any of its inflected forms. From the 1,867 resulting observations, 382 fulfilled our criterion of *translation direction* (Dutch as a source language, English as a target language). Each of the 382 sentences were manually annotated, meaning that the translation of *beginnen* was recorded for every sentence. For the example (1) below, 'take up' was annotated as the translation of *beginnen*:

(1) SOURCE: Zo vermeldde iemand bijvoorbeeld: "Ongeveer 80 procent van de afgestudeerden van onze kunstacademie zal een carrière *beginnen* in de creatieve industrie". [Someone mentioned for example: "About 80 percent of the graduates of our academy of arts will begin a career in the creative industry".]

TARGET: For example, in one case "Around 80 percent of graduates from our art school will *take up* careers in the creative industries". (dpc-vla-001920-nl, our emphasis)

From the 382 observations, 46 were disregarded for further analysis. We distinguish three reasons for such elimination. Two of them apply to all data retrieval and annotation tasks in our study, the third one is specific to the case of *beginnen* with English as a Language B.

1. The sentence alignment is erroneous. In this case, it is technically possible to look up the complete texts from which the aligned sentences were extracted and re-align the sentence correctly. However, we chose to disregard the erroneously aligned sentences out of practical considerations: re-alignment would have been too time-consuming.
2. The source language lexeme under consideration is not translated at all (or no translation equivalent can be indicated in a straightforward way). Observations where the lexeme under study remains untranslated in the target sentence, such as in the following example (2), are disregarded for further analysis:

[27] Beginnen$_{ENG}$ refers to the semantic mirroring initiated by the initial lexeme *beginnen* and with English as a language B.

(2) SOURCE: Ondernemers *begonnen* koortsachtig op zoek te gaan naar snoeiposten, [...]. [Entrepreneurs feverishly began to look for targets for cut backs]

TARGET: Company managers feverishly grasped to make savings, [...] (dpc-ing-002337-nl, our emphasis)

Although it would as such be interesting to examine why the inchoative aspect disappeared from the target sentence, this question is beyond the scope of the current study.

3. The third reason to eliminate an observation is when the lexeme *beginnen* is non-lexicalized in translation. This case is particularly relevant to the translation of Dutch *beginnen* into an English progressive structure (although similar translational situations are imaginable for this same verb and surely exist for other verbs, this is the only case encountered within our study of *beginnen* with English and French as languages B). Consider the following example 3:

(3) SOURCE: Terwijl de Europese Unie zich stilaan *begint* op te maken om 10 nieuwe lidstaten te verwelkomen, blijft de Europese economie een slappe bedoening.
[While the European Union begins gradually to prepare itself to welcome 10 new member states, the European economy remains a sluggish affair.]
TARGET: While the European Union *is* gradually prepar*ing* to welcome 10 new member states, the European economy remains in the doldrums. (dpc-ing-001896-nl, our emphasis)

In this particular example *zich opmaken* is translated by *to prepare* and *stilaan* is translated by *gradually*. The verb *beginnen* is not translated *lexically* here; instead its translation is couched in the structure 'to be + ing-form' applied to the verb *to prepare*. While we tried in an initial phase to preserve these observations by annotating the translation of *beginnen* as 'to be+ing-form', we ultimately had to refrain from integrating such structures into the further analysis. Its inclusion into the SMM++ was really like opening Pandora's box: it meant that the structure 'to be+ing-form' had to be queried from the corpus as a source language item in the *inverse T-image*. Although perfectly possible on the technical side, this soon appeared to be problematic: the inchoative aspect of the structure 'to be+ing-form' is often very subtle (see Smith 1997) and open for debate, as the following example (4) clarifies:

(4) SOURCE: But thanks to technological advances, plasma techniques *are playing* an ever greater role in our daily life: just think of fluorescent tubes and flat screen televisions, for example.

TARGET: Dankzij de technologische ontwikkeling duiken steeds meer plasmatoepassingen op in ons dagelijks leven. Denken we maar aan de tl-lampen of aan het vlakke plasmascherm van televisietoestellen. [Thanks to technological development, more and more plasma applications are popping up in our daily live. Think of striplighting or the flat plasma screen of television sets.] (dpc-arc-002037-en, our emphasis).

In example 4, the pattern ‘to be+ing-form’ could arguably be said to carry an inchoative aspect. The Dutch target sentence in fact even provides evidence for the inchoative aspect: the verb *to play* is not translated into *spelen*, which would have been a perfectly acceptable translation solution and even the readiest one (*to play a role*, *een rol spelen*). Instead, the translator selected the verb *opduiken* [to pop up, to turn up] which lexicalizes the inchoative aspect of the ‘to be+ing-form’ pattern of the English source sentence. The potential relevance of such an observation is of course indisputable (for instance, for researchers interested in grammaticalization and lexicalization patterns) but this example also shows that a whole other approach is needed for the annotation and analysis of this type of verb patterns in the source text with their corresponding items in the target text (be it another pattern or a single verb). The reason is that one should also envisage and annotate the translation of those patterns into still other patterns in the target language. It should be clear that this would increase the complexity of the application of the SMM++ considerably, reducing one of its advantages, i.e. the straightforward annotation of a source language lexical item and its translation (into a lexical item). The omission of translations into verb patterns (those observations in which a verb pattern is proposed as a translation for the lexeme under study) could be seen as a shortcoming of this study. If we want to further use the SMM++ in the future, this problem definitely needs a solution. If such complex annotations are not compatible with the proposed method, the SMM++ may be less suited to investigate the semantic structures of word categories other than nouns. However, in our first application of the SMM++, we reasonably limit the factors of complexity and disregard this type of verb patterns. In the case of *beginnen*, this can be done by disregarding translations into ‘to be+ing-form’.

The 336 remaining observations for the *first T-image* of beginnen$_{ENG}$ (listed in Table 5) consist of 44 different translations. From those 44 lexemes, 35 were observed less than 3 times. In other words, only 9 translations met the *frequency threshold* of 3 observations. Those 9 translations account for 292 of the total of 336 observations. In Table 5, the lexemes in bold meet the frequency threshold of 3 observations and are selected for further analysis.

Table 5 First T-image of beginnen$_{ENG}$ (raw frequencies)

	beginnen
already	1
as from	1
aspiring	1

beginning (adj)	2
beginning (n)	**3**
first of all	**3**
fundamental	1
initial	1
introduction	1
nascent	2
new	1
original	1
start (n)	**7**
start-up (n)	1
to adopt	1
to assume	1
to be rooted	1
to bear	1
to begin	**89**
to come	1
to commence	2
to develop	1
to embark	2
to emerge	1
to enter	2
to gain	1
to go ahead	1
to go into	1
to kick off	1
to launch	2
to let	1
to open	**5**
to result	1
to see	1
to set up	**3**
to start	**171**
to start off	2
to start out	**6**

to start up	**5**
to take up	2
to talk	1
to try	1
to undertake	2
young	1
TOTAL	336

Table 6 gives a summary the first step of the SMM++ retrieval task:

Table 6 First T-image of beginnen$_{\text{ENG}}$

Step of the SMM++	FIRST T-IMAGE
Source language	Dutch
Target language	English
Total queried observations	382
Total selected observations after discarding erroneous alignments and non-translated observations	336
Total different translations	44
Total selected observations after frequency threshold	292
Total selected different translations after frequency threshold	9
Source language lexeme(s)	beginnen
Selected target language lexemes	1. beginning (n) 2. first of all 3. start (n) 4. to begin 5. to open 6. to set up 7. to start 8. to start out 9. to start up

3.5.1.2 Inverse T-image of beginnen$_{\text{ENG}}$

The next step of the SMM++ consists in querying the 9 lexemes from the *first T-image* as source language lexemes in the DPC (all English sentences containing each of these 9 lexemes are queried, only those sentences where English is the source language and Dutch the target language are selected). For each observation, the translation back into Dutch of the lexeme is annotated, which leads to the summary in Table 7:

Table 7 Inverse T-image beginnen$_{ENG}$

Step of the SMM++	INVERSE T-IMAGE	
Source language	English	
Target language	Dutch	
Total queried observations	1217	
Total selected observations after discarding erroneous alignments and non-translated observations	1029	
Total different translations	148	
Total selected observations after frequency threshold and overlap	829	
Total selected different translations after frequency threshold and overlap	24	
Source language lexeme(s)	1. 2. 3. 4. 5. 6. 7. 8. 9.	beginning (n) first of all start (n) to begin to open to set up to start to start out to start up
Selected target language lexemes	1. aanvang 2. (allereerst) 3. begin 4. beginnen 5. eerst 6. gaan 7. inzetten 8. komen 9. krijgen 10. maken 11. ontstaan	13. opening 14. oprichten 15. opstarten 16. opzetten 17. sinds 18. start 19. start- 20. starten 21. steeds meer 22. van start gaan 23. vanaf

	12. openen	24. worden

3.5.1.3 Second T-image of beginnen$_{\text{ENG}}$

The following step of the SMM++ consists in querying the 24 lexemes from the *inverse T-image* as Dutch source language lexemes in the DPC. For each selected observation, the translation of the source lexeme back into English is annotated. As mentioned in 3.4.3, these data are selected according to an additional restriction, i.e. their translations have to be identical to one of the source language lexemes of the *inverse T-image*. In practice, there are two implications: first, the total number of selected observations is 17 times smaller than the (enormous) total number of queried observations[28], and second, one source language lexeme *allereerst* had to be discarded because its back-translations into English did not match any of the 9 selected target language lexemes (a problem most probably due to corpus size). This final results of the mirroring are summarized in the following Table 8:

Table 8 Second T-image of beginnen$_{\text{ENG}}$

Step of the SMM++	SECOND T-IMAGE	
Source language	Dutch	
Target language	English	
Total queried observations	20869	
Total selected observations after restriction rule	(1182) 1117[29]	
Source language lexeme(s)	1. aanvang 2. (allereerst) 3. begin 4. beginnen 5. eerst 6. gaan 7. inzetten	13. opening 14. oprichten 15. opstarten 16. opzetten 17. sinds 18. start 19. start-

[28] In order to cope with the large amount of observations, we carried out a preliminary statistical word alignment using GIZA++. Every statically word-aligned observation was subsequently manually verified.

[29] The number between brackets indicates the total number of selected observations in the *second T-image* of beginnen$_{\text{ENG}}$, the second number refers to the total number of observations for the *second T-image* of beginnen$_{\text{ENG}}$ after the selection of only those lexemes which are also members of the *second T-image* of beginnen$_{\text{FR}}$. See section 3.4.3.

	8. komen 9. krijgen 10. maken 11. ontstaan 12. openen	20. starten 21. steeds meer 22. van start gaan 23. vanaf 24. worden
Target language lexemes	1. 2. 3. 4. 5. 6. 7. 8. 9.	beginning (n) first of all start (n) to begin to open to set up to start to start out to start up

3.5.2 SMM++ of beginnen$_{FR}$[30]

The retrieval task of the SMM++ was also carried out with French as a pivot language. The tables below summarize the results of each steps of the mirroring.

3.5.2.1 First T-image of beginnen$_{FR}$

Table 9 summarizes the information of the *first T-image* of beginnen$_{FR}$:

Table 9 First T-image of beginnen$_{FR}$

Step of the SMM++	FIRST T-IMAGE
Source language	Dutch
Target language	French
Total queried observations	472
Total selected observations after discarding erroneous alignments and non-translated observations	398

[30] Beginnen$_{FR}$ refers to the semantic mirroring initiated by the initial lexeme *beginnen* and with French as a language B.

Total different translations	75	
Total selected observations after frequency threshold	332	
Total selected different translations after frequency threshold	19	
Source language lexeme(s)	beginnen	
Selected target language lexemes	1. à partir de 2. commencer 3. d'abord 4. début 5. débutant (adj) 6. débutant (n) 7. débuter 8. démarrer 9. entamer 10. entreprendre	11. entrer 12. lancer 13. lancer, se 14. mettre, se 15. ouvrir 16. partir 17. prendre cours 18. (prendre son depart) 19. recommencer

3.5.2.2 Inverse T-image of beginnen$_{\text{FR}}$

Table 10 summarizes the results of the *inverse T-image* of beginnen$_{\text{FR}}$:

Table 10 Inverse T-image of beginnen$_{\text{FR}}$

Step of the SMM++	INVERSE T-IMAGE
Source language	French
Target language	Dutch
Total queried observations	2409
Total selected observations after discarding erroneous alignments and non-translated observations	1706
Total different translations	339
Total selected observations after frequency threshold and overlap	1179
Total selected different translations after frequency threshold and overlap	39

Source language lexeme(s)	1. à partir de 2. commencer 3. d'abord 4. début 5. débutant (adj) 6. débutant (n) 7. débuter 8. démarrer 9. entamer	10. entreprendre 11. entrer 12. lancer 13. lancer, se 14. mettre, se 15. ouvrir 16. partir 17. prendre cours 18. recommencer
Selected target language lexemes	1. aanvang 2. aanvangen 3. aanvankelijk 4. aanvatten 5. begin 6. begin- 7. beginnen 8. belanden 9. doen 10. een aanvang nemen 11. eerst 12. gaan 13. in werking treden 14. ingaan 15. komen 16. krijgen 17. lanceren 18. maken 19. nemen 20. ondernemen	21. ontstaan 22. ontwikkelen 23. op basis van 24. openen 25. oprichten 26. opstarten 27. opzetten 28. sinds 29. sluiten 30. start 31. starten 32. storten, zich 33. ten eerste 34. uitgaan van 35. van start gaan 36. vanaf 37. vanuit 38. vertrekken 39. worden

With regard to the *inverse T-image* of beginnen$_{FR}$, there are two points which require our further attention: the first one is the lexeme *prendre son départ* and the second one relates to the proportion of selected data versus the total of queried data.

Prendre son départ

The lexeme *prendre son départ* was initially selected as one of the source language lexemes of the *inverse T-image* of beginnen$_{FR}$ (since it met the condition of frequency threshold of 3 observations in the *first T-image*). However, no observations were found with *prendre son départ* as a French source language expression. Two explanations are plausible. First, on closer analysis, all observations of the *first T-image* which rendered *prendre son départ* as a translation, appeared to stem from two documents (dpc-wst-000014-fr and dpc-wst-000071-fr) which were translated by the same two translators and released by the same text provider. This could suggest that we were dealing with an (quasi-)idiosyncratic expression from the two translators. However, the two documents (dpc-wst-000014-fr and dpc-wst-000071-fr) also share the same subject: they describe walks/walking trails for tourists. This seems in fact to be a typical context in which the expression *prendre son départ* appears, as the following examples (5 and 6) from the FrWaC[31] corpus confirm:

(5) Le parcours vallonné *prend son départ* au lotissement de Saint Paul près de la chapelle , traverse le Pont de Reynès et monte au travers de la montagne jusqu' au village . [The hilly path *starts from* the townsite of Saint Paul's near the chapel, crosses the Reynès bridge and goes up accross the mountain to the village.] (corpus position 94673986, our emphasis)

(6) Quant au chemin de fer touristique du Tarn , il *prend son départ* à l' ancienne station des Tramways à vapeur du Tarn au centre de Saint-Lieux . [As far as the tourist railway of the Tarn concerns, it *starts off* in the old station for steam trams of the Tarn in the centre of Saint-Lieux.] (corpus position 269689, our emphasis)

Other contexts in which *prendre son départ* can be used are more philosophical in nature, as the following example 7 illustrates:

(7) Le propos de Laplanche *prend son départ* , en effet , de l' idée qu' éros-liaison oeuvre en tant que tel « dans un sens narcissique » , puisqu' il tend , dit -il , à « faire de l' un » (Lacan) . [Laplanches comment indeed stems from the idea that the eros connection is as such as work "in a narcissistic way", because it tends, so he says, to "the becoming of one" (Lacan)]. (corpus position 60635066, our emphasis)

These examples show that the lack of observations for *prendre son départ* as a source language lexeme is not so much due to idiosyncratic language use, but rather to data sparseness in the DPC. Although this makes it clear that *prendre son départ* can be considered as an accepted expression of inchoativity in French, its use is restricted to very specific contexts which the DPC does not provide. One could argue that the above offers substantial insights to include *prendre son départ* for further analysis. Although this is true of course, we cannot do so for obvious pragmatic reasons: the DPC simply does not provide any data with *prendre son départ* as a source language lexeme, so

[31] FrWac is a 1.6 billion word, web-derived corpus (Ferraresi et al. 2010) which we consulted here for reference.

further mirroring cannot be carried out for this verbal expression. Furthermore, we do not have a larger, parallel and comparable corpus for Dutch, French and English which could provide additional observations for *prendre son depart* at our disposal.

A second point which can be made here is that the final selection of data for beginnen$_{FR}$ is proportionally smaller than the selection for beginnen$_{ENG}$ – a little over 70%, compared to more than 80% for beginnen$_{ENG}$. This is due to a higher ratio of erroneous alignments, but appears to be often the result of an omission in the translation. Translating by omission is one of the strategies indicated by Baker (1992, 40). It is an interesting phenomenon which should not be neglected and from which interesting findings can ensue. In our study, for example, no translation into Dutch could be formally indicated in 59 out of 226 observations for the French adverb *d'abord* (over 26% of the cases). By contrast, its English equivalent *first of all* is translated into Dutch in 17 out of 18 observations. Hence, it appears that translators more easily omit French *d'abord* when translating into Dutch than English *first of all* when translating into the same language. Interestingly, such contrastive comparisons of translation by omission can reveal diverging patterns of translational behavior for different languages and different parts of speech. Unfortunately, we have to discard the observations of translation by omission (as was already explicated in section 3.5.1.1. of this chapter) as zero translations cannot be selected and retrieved as a source language lexeme in the next step of the SMM++.

3.5.2.3 Second T-image of beginnen$_{FR}$

Table 11 recapitulates the results of the *second T-image* of beginnen$_{FR}$:

Table 11 Second T-image of beginnen$_{FR}$

Step of the SMM++	SECOND T-IMAGE
Source language	Dutch
Target language	French
Total queried observations	26317
Total selected observations after restriction rule	(1822) 1490[32]

[32] The number between brackets indicates the total number of selected observations in the *second T-image* of beginnen$_{FR}$, the second number refers to the total number of observations for the *second T-image* of beginnen$_{FR}$ after the selection of only those lexemes which are also members of the *second T-image* of beginnen$_{ENG}$. See section 3.4.3.

Source language lexeme(s)	1. aanvang 2. aanvangen 3. aanvankelijk 4. aanvatten 5. begin 6. begin- 7. beginnen 8. doen 9. een aanvang nemen 10. eerst 11. gaan 12. in werking treden 13. ingaan 14. komen 15. krijgen 16. lanceren 17. maken 18. nemen	19. ondernemen 20. ontstaan 21. op basis van 22. openen 23. oprichten 24. opstarten 25. opzetten 26. sinds 27. sluiten 28. start 29. starten 30. ten eerste 31. uitgaan van 32. van start gaan 33. vanaf 34. vanuit 35. vertrekken 36. worden
Target language lexemes	1. à partir de 2. commencer 3. d'abord 4. début 5. débutant (adj) 6. débutant (n) 7. débuter 8. démarrer 9. entamer	10. entreprendre 11. entrer 12. lancer 13. lancer, se 14. mettre, se 15. ouvrir 16. partir 17. prendre cours 18. recommencer

A few points need to be made for the *second T-image* of beginnen$_{FR}$. Firstly, the Dutch source language lexemes *belanden*, *ontwikkelen* and *zich storten* are excluded for further analysis because none of their translations matched one of the French target language lexemes. The following paragraphs try to explain why these lexemes were selected in the first place and subsequently discarded.

Belanden [to end up at]

In the *inverse T-image*, *belanden* [to end up at] was annotated three times as a translation of *entrer* [to enter], and once as a translation of *début* in the expression e*ffectuer ses* *débuts* [making your debut]. Further analysis revealed that those three observations (where *entrer* was translated by *belanden*) were all attested in the same document (dpc-lan-001629-fr), translated by the same translator and treating the same subject, i.e., to enter in politics. *Belanden* was filtered out by the restriction rule of the *second T-image*: none of its translations into French match the source language lexemes of the *first T-image*. This indicates that the inchoative aspectual meaning of *belanden* is (very) rare, to the point that it is attested in none of the 29 observations of the verb. Instead, *belanden* is rather translated by *arriver* [to arrive], *atterrir* [to land] or *se retrouver* [to meet].

Ontwikkelen [to develop]

As for *ontwikkelen* [to develop], we see that in the *inverse T-image* it was three times annotated as a translation of *lancer* [to launch] and once as a translation of *entrer* [to enter]. Close inspection of the three observations for *lancer –ontwikkelen* shows that two of them (examples 8 and 9) were amenable to a different annotation:

(8) SOURCE: A noter que nous sommes en train de *lancer* et développer des outils pour faire davantage vivre cette communauté d'amoureux de musique.
[Note that we are launching and developing a number of tools to bring this music-loving community even more to live.]
TARGET: We zijn trouwens volop bezig tools te ontwikkelen om deze community van muziekliefhebbers meer animo te geven.
[We are by the way very busy developing tools to bring more gusto in this community of music lovers.] (dpc-rou-003216-fr, our emphasis)

(9) SOURCE: La marque de jeans Diesel a, par exemple, *lancé* un concours aux membres de Facebook, par le biais d'une application, baptisée 'comment vivez-vous avec votre Diesel?'.
[The jeans brand Diesel has, for example, launched a contest for its Facebook members, via an application baptized ‘how do you live with your Diesel?’.]
TARGET: Zo ontwikkelde het jeansmerk Diesel een applicatie voor een wedstrijd onder Facebookleden, 'hoe leef jij met je Diesel?'.
[The jeans brand Diesel developed an application for a contest amongst Facebookmembers, ‘how do you live with your Diesel?’.]

The verb *ontwikkelen* in example 8 was annotated as the translation of *lancer*. One could indeed argue that, since only one verb is retained in Dutch, i.e. *ontwikkelen*, this verb embodies both *lancer* and *développer*. Alternatively, it could also be claimed that the translation of *lancer* is not *ontwikkelen* but a zero translation.

In example 9, the verb *ontwikkelen* was annotated as the translation of *lancer*. Close inspection of source and target sentences in this example shows that the target sentence is open for two different interpretations. In the first case, *ontwikkelen* has in fact not been translated at all: whereas the French source language sentence reads ‘a contest was launched via an application’, the Dutch translation by contrast reads ‘an application was developed for a contest', omitting the verb *lancer* [to launch] and adding *ontwikkelen* [to develop]. The other interpretation is that *lancer* also refers to *application* in the French source language sentence so *ontwikkelen* can be considered as its correctly annotated translation. This example shows how difficult the annotation task sometimes can be[33]. However, because of the restrictions on the *second T-image* (see 3.4.3), *ontwikkelen* has been excluded from the analysis.

Zich storten [throw oneself, plunge]

Finally, the reflexive verb *zich storten* was observed 3 times as a translation of *se lancer* [to launch oneself] and once of *se mettre* [to begin]; all observations stem from different texts, translated by different translators; the annotation of the translations is furthermore unequivocal, so that *zich storten* was initially selected. However, *zich storten* did not meet the restrictions for the *second T-image*, so it was excluded from the analysis. As a consequence, this can be considered as a symptom of (lack of) corpus size: given the success rate of *zich storten* in the inverse T-image, a larger corpus would certainly have included it in the analysis (although it would probably not have shown up as a prototypical expression of inchoativity). This third example therefore shows that larger corpora are necessary for the inclusion of less prototypically used lexemes.

3.5.2.4 Final selection of candidate lexemes

Tables 8 and 11 (summarizing the *second T-images* of beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$) respectively contain two numbers for the final total number of observations. The number between brackets represents the total number of observations when carrying out the procedure as has been described above. The second (smaller) number involves one last practical issue which needs to be resolved for the purpose of the statistical analyses and visual comparisons of *all* the retrieved data sets. In order to be able to compare the *second T-image* of beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$ with the *inverse T-images*, we have to select the *common* lexemes of the *second T-images* of beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$. As the summaries of beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$ show, an SMM++ which is carried out with a same initial lexeme but with different languages B does indeed not result into identical sets of Dutch lexemes (although the majority of the Dutch lexemes yielded in the *inverse T-image* are common for beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$). In total, 17 lexemes have been independently selected by both the mirroring of beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$.

[33] The reliability of the annotation was verified on the basis of a calculated inter-annotator agreement using Cohen's kappa statistic. An average kappa score of 0.79 was obtained for a random sample of 472 observations for the *first T-image* of beginnen$_{\text{FR}}$. This is considered as a reliable agreement (Carletta 1996).

These 17 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], *vanaf* [as from], *worden* [to become][34].

Technically speaking, this final step is not indispensable: it is possible to create visualizations of the complete sets of lexemes reproduced in tables 8 and 11, but renouncing this final restriction of the data set would have two implications. Firstly, the data of the *second T-images* of beginnen$_{ENG}$ and beginnen$_{FR}$ could not be merged, meaning that the data set of SourceDutch would be based on either beginnen$_{ENG}$ or beginnen$_{FR}$ – which would consequently take away the previously established 'safety mechanism' of merging the two sets in order to eliminate possible target language effects (see section 3.4.3). Secondly, the sets of lexemes whose visualizations we aim to compare would consist of different lexemes for either set, complicating the comparison of those visualizations. Taking all this into account, and conscious about the possible consequences of restricting our data sets with respect to their informativity, we opt for the security of comparing likes with likes in our final visualization step by selecting only those lexemes which the SMM++ of beginnen$_{FR}$ and beginnen$_{ENG}$ have in common.

[34] 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, aanvatten, begin-, doen, lanceren, maken, nemen, sinds, start-*.

3.6 Statistical visualization

After the application of the newly developed SMM++ for the retrieval of candidate-lexemes, we will now take the final methodological step of statistically analyzing the data. A visual exploration of the data seems to be the best option for this study since no clear hypotheses can be formulated yet for semantic differences in translation.

One of the main adaptations to the SMM proposed in the previous sections is the integration of frequency information into the rationale. The result of the SMM++ can be resumed in different data matrices which contain this frequency information. In this section, we will select an appropriate statistical visualization method which takes into account this newly gained information. Parallel to the 'natural' step in distributionalist semantics towards statistical methods, we want to propose a statistical visualization technique of the frequency tables obtained via the SMM++.

In order to select such an appropriate technique, we first need to carefully analyze the type of data we are dealing with. In the previous sections, we have shown that the data resulting from the SMM++ are resumed in frequency tables (or matrices).[35] The matrices list observations in the rows; the columns are considered as the attributes or properties of those rows (Baayen 2008, 118). Our aim is to discover structure in the data sets by grouping the observations according to their properties. One way to do so is by representing the lexemes in a spatial map. For frequency tables, this can be done with *correspondence analysis* (Greenacre 2007). We will carry out a first visual exploration of the data on the basis of *correspondence analysis* in section 3.6.1. A visualization of CA represents the first two latent dimensions of the CA. We will see, however, that for our data, the first two latent dimensions represent less than the established threshold of 80% of the inertia (although they do still represent 40 to 60%). It will become clear that – due to the subtlety of the semantic field we are describing – the delimitation of clearly distinct clusters in the CA is difficult and that the relations between the lexemes in the delimited clusters also remain unclear.

In order to overcome the above mentioned problems, we choose to combine Correspondence Analysis with Hierarchical Cluster Analysis (section 3.6.2). HCA is an unsupervised clustering technique, meaning that "the result of the clustering only depends on natural divisions in the data" (Manning & Schütze 1999, 498). More specifically, we will carry out a Hierarchical Agglomerative Clustering on the output of the CA. This means that the obtained coordinates of the CA will be used as an input for our HAC, a procedure which allows us to filter out noisy data. Each of the remaining

[35] The contingency tables for all data sets can be found in appendices 1 to 6.

sub-sections of 3.6.2 is concerned with a particular choice which needs to be made before the HAC can be carried out. In sections 3.6.2.1 and 3.6.2.2, we will put forward our choice of a particular (dis)similarity measure (Euclidean) and clustering algorithm (Ward's) respectively. In section 3.6.2.3, we will explain our procedure to determine the number of clusters and we will propose a validation procedure for the number of clusters. Finally, section 3.6.2.4 includes a comparison of the applied procedure (Euclidean distance, Ward's Minium Variance Method, HCA on the output of CA) to other, alternative procedures which include the use of a distinct distance measure (Canberra), clustering algorithms (average and complete linkage), and data input for the HCA (raw data and output of a LSA).

In section 3.6.3 we will propose a number of statistical tools to reveal the prototype-based organization of the clusters in a dendrogram and of the lexemes within each cluster. These measures will be used to investigate semantic *levelling out*. In section 3.6.4, we also put forward two additional analyses which can help us to interpret the influence of a specific source language on the translated semantic fields: the visualization of the SourceField of the language B (for the investigation of semasiological *shining through*) and Multiple Correspondence Analysis on the Burt tables of the TransDutch fields (for the investigation of onomasiological *shining through*). All our analyses were 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 used the svs-package (Plevoets 2015) which contains "various tools for semantic vector spaces" for a number of analyses. We used the function fast_sca() from the svs-package to carry out the CA. While the same result could indeed be obtained via the existing function ca(), the svs-function fast_sca() is especially designed to further use the resultant coordinates as the input for an additional analysis (in our case, we will use the output of a CA as the input for a HAC).

3.6.1 Correspondence Analysis

Correspondence Analysis, 'a special case of multidimensional scaling' (Baayen 2008, 136), seems a good candidate technique to map our frequency tables in a low-dimensional space:

> Correspondence Analysis (CA) – a method of displaying the rows and columns of a table as points in a spatial map, with a specific geometric interpretation of the positions of the points as a means of interpreting the similarities and differences between rows, the similarities and differences between columns and the association between rows and columns (Greenacre 2007, 264).

Essentially, CA works as follows: given a fictitious data matrix in Table 12, the objective is to display the Dutch lexemes in the rows and the language B lexemes (in this example French lexemes) in the columns as points in a spatial map.

Table 12 Fictitious data matrix for CA

	commencer	débuter	début	départ
beginnen	7	5	4	3
starten	5	4	2	2
aanvangen	0	3	2	0
aanvatten	2	0	1	1
v start gaan	3	5	0	0

The initial map has as many dimensions as there are columns in the data matrix (Figure 16).

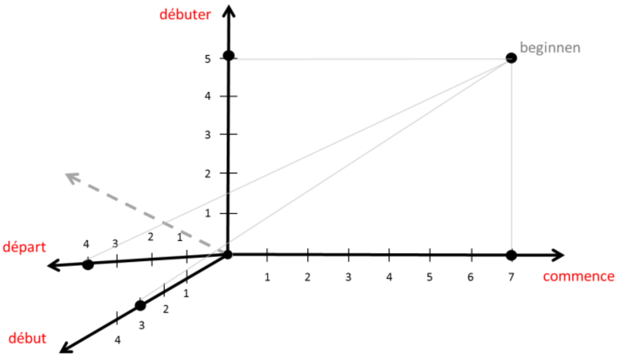


Figure 16 Spatial map with n dimensions for *beginnen*

Now, in order to be able to visually present the specific geographic position of each of the Dutch lexemes in the rows, its position in the *n*-dimensional space is reduced to a two-dimensional space (Figure 17).

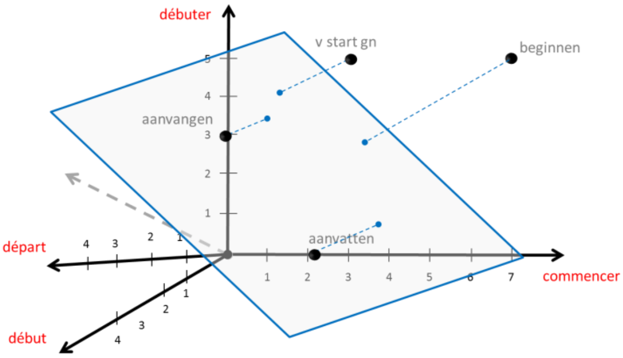


Figure 17 Reduction to a two dimensional space for all rows

All five Dutch lexemes can be represented as points in this space. Next, the best fitting two-dimensional space is computed (Figure 18). Because this two-dimensional map captures the original high-dimensional data cloud as much as possible, it is true that "the larger the distance between two rows, the further these two rows should be apart in the map for rows" (Baayen 2008, 129). Consequently, the positions of the lexemes and the distances between the plotted lexemes represent the similarities and differences between the lexemes. The same computation is repeated for the columns of the frequency table and the simultaneous representation of the row map and the column map results in a so-called bi-plot (representing the scatterplot of the row map and the scatterplot of the column map simultaneously).

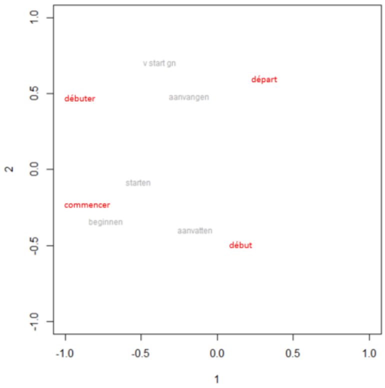


Figure 18 Bi-plot for fictitious data matrix in Table 12

When we now apply CA to our own data, we obtain a first visualization via CA of the SourceDutch field of *beginnen* (Figure 19).

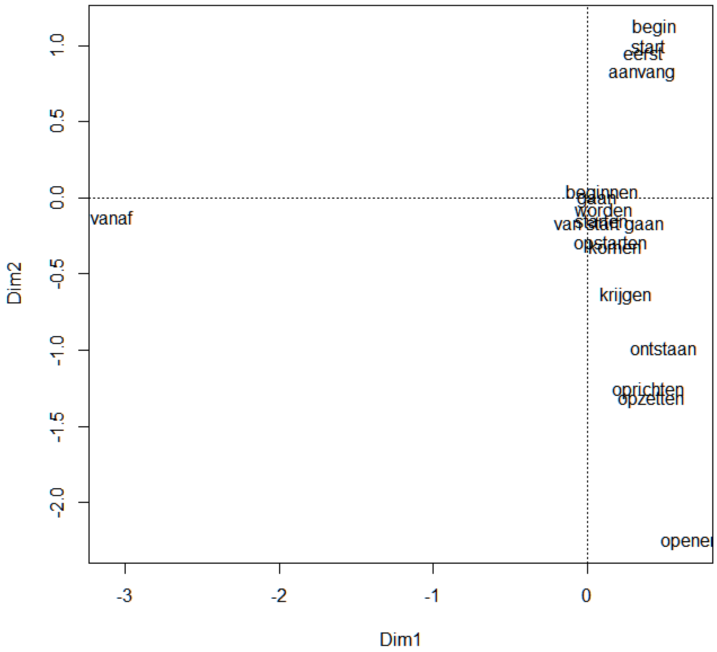


Figure 19 First Correspondence Analysis of SourceDutch field for *beginnen*

What is immediately striking is the outlying position of *vanaf.* Although the selection of lexemes has been done through a careful technique (see the previous sections) we decide to exclude *vanaf* from all data sets. Looking back at the frequency tables (the *second T-images* of beginnen$_{\text{ENG}}$ and beginnen$_{\text{FR}}$, see appendices 5 and 6) for SourceDutch, we indeed see that *vanaf* has an "unusual profile" (Greenacre 2007, 92): *vanaf* is related to a single French target lexeme, i.e. *à partir de*. In the *second T-image* of beginnen$_{\text{FR}}$, we also see that the relative weight of *vanaf* is rather high (0.1505792; representing thus 15 % of the total number of observations) and contributing to a 0.1953608 – over 19% – rise of the total inertia[36] of the data matrix when compared to the same data matrix without *vanaf.* The conclusion is that the variation of the first dimension is solely accounted for by *vanaf*. Greenacre (2007, 92) indeed warns for the fact that outliers can "start dominate a map so much that the more interesting contrasts between the more frequently occurring categories are completely masked". The data points in the plot without *vanaf* (Figure 26) are indeed more spread out in the two-dimensional space, which will facilitate the interpretation. Based on the above, we decide to remove *vanaf* from all our data sets.

[36] "1. The (total) inertia of a table quantifies *how much variation* is present in the set of row profiles or in the set of column profiles" [...] 3. CA is performed with the objective of accounting for a maximum amount of inertia along the first axis. The second axis accounts for a maximum of the remaining inertia, and so on. [...]" (Greenacre 2007, 88, our emphasis).

Before we analyze a visualization via CA, we first need to assess the degree of representativeness of the plots with respect to the total variation in each of the data sets. The measure for variation in a frequency table is the inertia (Greenacre 2007). The distribution of inertia over the latent dimensions of the CA can be visualized in a so-called scree plot: the bars show how much of the total variation is associated with each dimension. Consequently, the scree plot indicates how many dimensions are needed to reach a threshold, e.g. 80%. If we take a look at the scree plots for SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$, we see that five dimensions are required for SourceDutch (Figures 20 and 21), three dimensions for TransDutch$_{ENG}$ (Figures 22 and 23) and four dimensions for TransDutch$_{FR}$ (Figures 24 and 25) for 80%. This presents a practical problem, however, as 4 or 5 dimensional plots are not easily visualized. Although we are aware that a visualization via CA for SourceDutch only represents around 40% of the inertia, we will use the visualization in Figure 26 as a first, exploratory analysis of the field of SourceDutch.

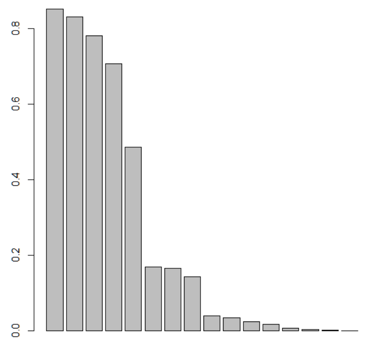


Figure 20 Scree plot for SourceDutch

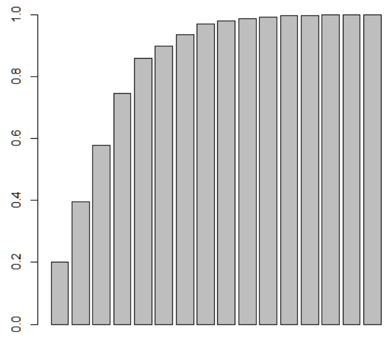


Figure 21 Cumulative scree plot for SourceDutch

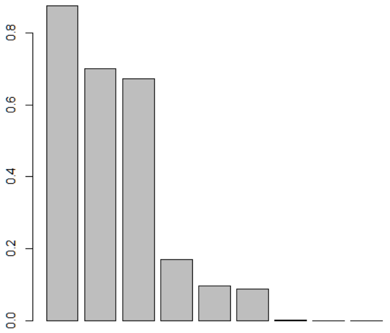


Figure 22 Scree plot for TransDutch$_{\text{ENG}}$

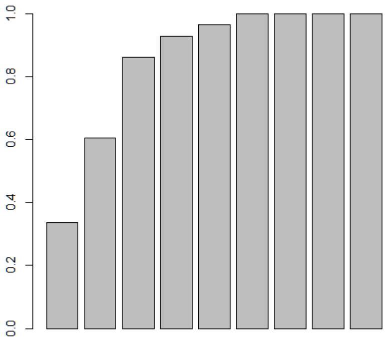


Figure 23 Cumulative scree plot TransDutch$_{\text{ENG}}$

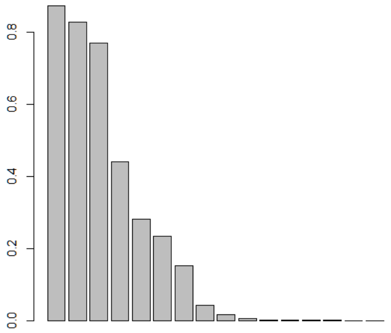


Figure 24 Scree plot for TransDutch$_{\text{FR}}$

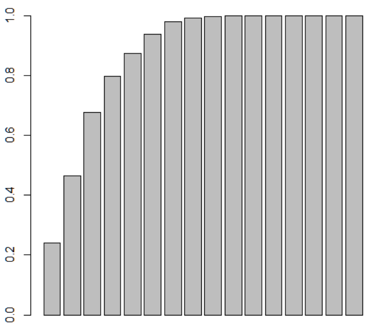


Figure 25 Cumulative scree plot for TransDutch$_{\text{FR}}$

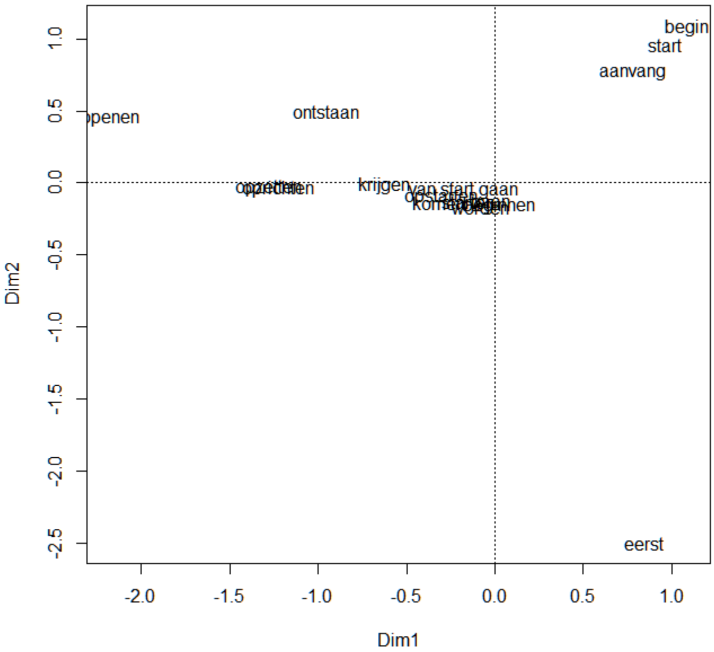


Figure 26 Correspondence Analysis of SourceDutch field for *beginnen* without *vanaf*

In Figure 26, we observe one large central cluster, situated around the origin (the 'zero-point') of the plot which contains, amongst other lexemes, the initial lexeme *beginnen*. We consequently interpret this central cluster as the prototypical center, consisting of lexemes with the basic meaning of the inchoative category, viz. "start of a general process". In the upper right corner, a second cluster contains *aanvang* [commencement], *start* [start] and *begin* [beginning]; all three lexemes are nouns, where *start* and *begin* are the nominal derivatives of *beginnen* and *starten* (which belong to the cluster considered as the prototypical center). The third lexeme *aanvang* then, is the more formal[37] counterpart of *begin* and *start*. In the lower right corner, we see *eerst* [firstly], which holds a somewhat outlying position. This outlying position can be explained by the fact that the translations which determine its position (*d'abord* and *firstly*) are almost exclusively used as translations of *eerst*. We furthermore see *oprichten* [to establish] and *opzetten* [to set up] clustering together. In the lexical database Cornetto (Vossen et al. 2008; 2013) *oprichten* is defined as *opzetten* and both verbs are indicated to refer to inchoative situations involving a project, a business, a company, etc. In other words, our CA confirms the strong relation between the two lexemes. Finally, *ontstaan* [to come into being] and *openen* [to open] occupy a somewhat unclear position between the center and

[37] In order to underpin the assertions we present with respect to the pragmatics or semantics of a given lexeme, we rely on information retrieved in the lexical database Cornetto (Vossen et al. 2013).

periphery of the graph. On the basis of the CA, we discern three different clusters: one central cluster (considered as the one with the most prototypical expressions of inchoativity); one cluster containing the nominal derivatives of *beginnen* and *starten* plus *aanvang*, a small third cluster with the near-synonymous verbs *oprichten* and *opzetten.* It is not entirely clear whether *ontstaan* and *openen* could be considered as one cluster, or whether they should be considered as two separate, singleton clusters.

Due to the subtlety of the semantic field that we are describing, the delimitation of clearly distinct clusters can thus appear difficult. A drawback of CA moreover is that it does not allow us to further analyze the central cluster: the visualization only suggests that the lexemes within this cluster are closely related, but the exact relations remain unclear.

Conclusively, we could make the following observations on the basis of this preliminary CA. Firstly, we were able to detect and remove an outlying data point which was distorting the overall interpretation of the data (*vanaf*). Secondly, the scree plots for SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ showed that for our data, more than 2 dimensions are required to accurately represent the distribution of the inertia over the latent dimensions of the CA. This represents a practical problem with respect to our visualization purposes. Thirdly, a first exploration of the SourceDutch field on the basis of the CA allowed us to formulate some preliminary insights into the semantic field. However, we also found that the delimitation of clearly distinct clusters appeared difficult and that we could not further examine the exact relations between the lexemes in the central cluster. We therefore decide to use Hierarchical cluster Analysis for the visualization of the semantic fields of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$. As we already announced in the introduction of this section, the HCA will be carried out on the output of a CA, a procedure on which will be further elaborated in the next section.

3.6.2 Hierarchical Cluster Analysis

Hierarchical Cluster Analysis (HCA) is "a collection of different algorithms that *puts objects into clusters* according to well-defined similarity rules" and is "mostly used when we do not have any a priori hypotheses" (Divjak & Fieller, 2014, 406, our emphasis). In this section, we will first describe which type of cluster analysis seems to be the best choice for our study. In addition, as every cluster analysis is crucially dependent on both a particular similarity measure and clustering algorithm, we will elaborate on these measures in section 3.6.2.1 and section 3.6.2.2 respectively. Next, we will explain the procedure for determining the number of clusters for which we will rely on the `R` package `pvclust` (Suzuki & Shimodaira 2006) (section 3.6.2.3). Finally, we propose to validate the combined choice of a particular similarity measure and clustering algorithm and for the number of clusters in the cluster solution (section 3.6.3.4)[38].

Just as semantic spaces are customary in computational semantics, in (cognitive) linguistics, "[c]luster analyses have been used to determine the similarity of intraword senses or the degree of granularity exhibited by polysemous word senses (cf. Miller 1971; Sandra and Rice 1995; Rice 1996)" (Gries 2006b, 81). The method has also been extensively used by Gries and Divjak (see for example: Divjak 2006, Divjak 2010, Divjak & Fieller 2014, Divjak & Gries 2006, 2009, Gries 2006b, Gries & Divjak 2009, Gries & Otani 2010, Dehors & Gries 2014). The reasons for HCA's popularity are summarized by Divjak:

> Cluster analysis is one of the basic exploratory techniques that are often applied in analyzing large data sets. This statistical method helps organize observed data into meaningful structures: it finds similarities between elements and groups similar elements together. These groupings, in turn, assist in understanding relationships that might exist among these elements. In other words: cluster analysis finds the most optimal solution and organizes an enormous number of data in substructures that facilitate comparison of the (elements in the) structures to each other (Divjak, 2010, 129-130).

HCA is not a single technique, but covers "a family of techniques for clustering data and displaying them in a tree-like format" (Baayen 2008, 138). In Statistical NLP, HCA has two main uses: *exploratory data analysis* (EDA) on the one hand and *generalization* on the other hand (Manning & Schütze 1999, 497). The tree-like format in which the result of a clustering algorithm can be visually represented is called a dendrogram:

[38] Exhaustive overviews of the existing clustering techniques can be found in Manning & Schütze (1999, 495-528), Baayen (2008, 138-148), Everitt et al. (2011, 71-110), Gries (2013, 336-349) and Divjak & Fieller (2014).

> a branching diagram where the apparent similarity between nodes at the bottom is shown by the height of the connection which joins them. Each node in the tree represents a cluster that was created by merging two child nodes. [...] The "height" of the node corresponds to the decreasing similarity of the two clusters that are being merged (Manning & Schütze, 495).

In order to maintain terminological clarity, we propose the following terminology (visualized in Figure 27), which is to a large extent based on Everitt et al. (2011, 89). A *node* can refer to either an *internal node*, a *sub-node* (an internal node within one delimited cluster) or a *terminal node* (also called a *leave*). The *heights* of the *edges* can be read off from the dendrogram. The line perpendicular to the edges in the tree is called the *root*. Finally, we will call the names printed at the extremities of every terminal node *lexemes* or *lexical items* (which is an immediate adaptation of the terminology to the type of data we are working with) instead of the term *label* proposed by Everitt et al (2011).

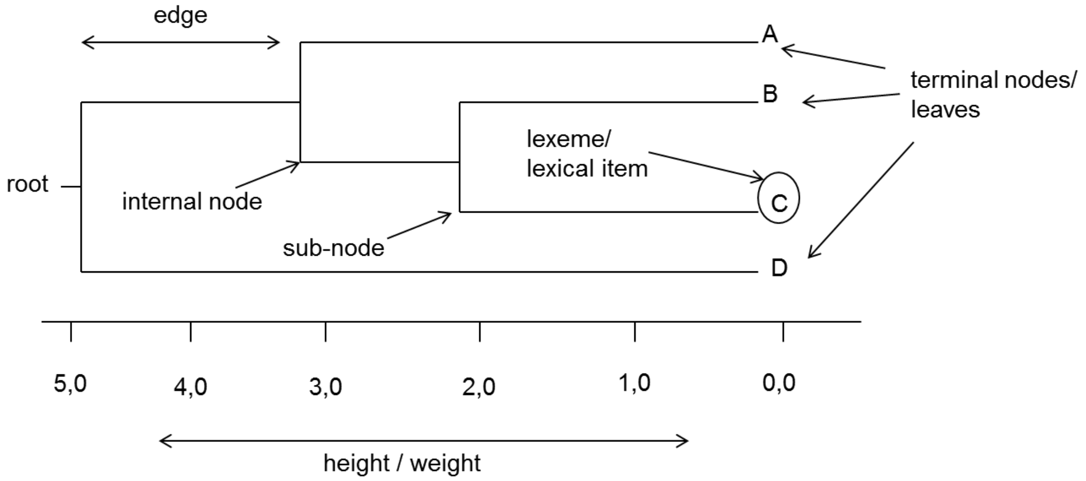


Figure 27 Terminological description of a dendrogram (adapted from Everitt et al. 2011, 89)

HCA comes in two flavors: the *tree* can be constructed either top-down or bottom-up. The first method is called *divisive clustering* where "one starts with all the objects and divides them into groups so as to maximize within-group similarity" (Manning & Schütze 1999, 501). The second method is called *agglomerative clustering* which works "by starting with the individual objects and grouping the most similar ones (Ibid., 500-501)". *Divisive clustering* – also called *partitioning* – is known to have difficulties in finding "optimal divisions for smaller clusters" and appears to be better at finding a few large clusters (Baayen 2008, 138). This can be verified by our visualized result (Figure 28), which shows a so-called *chaining effect* when applying divisive clustering for TransDutch$_{ENG}$. This means that the cluster tree displays "a chain of large similarities without taking into account the global context" (Manning & Schütze 1999, 504). As Manning and Schütze argue, cluster analysis is normally based on "the assumption that 'tight' clusters are better than 'straggly' clusters", and that this in turn "reflects an intuition that a cluster is a group of objects centered around a central point, and so compact clusters are to be preferred" (p.506). In particular, this corresponds to "a model like the Gaussian distribution"(Ibid.). Although Manning and Schütze stress that this is "only one possible underlying model of what a good cluster is", and that a good clustering should rely on prior knowledge or a model of the data, 'elongated clusters' due to a *chaining effect* are usually disfavored to sphere-shaped clusters (Ibid.). Because we will interpret dendrograms as semantic fields of *beginnen*, organized in a prototype-based manner - with the different clusters representing the meaning differentiations of the lexeme under study – we will prefer a clustering which indeed reflects our intuition that the clusters are centered around a central point and avoids large, elongated clusters caused by a *chaining effect*.

In summary, we agree with Everitt et al. (2011, 92) that the *chaining effect* is a symptom of distortion through "space contraction" where "dissimilar objects are drawn into the same cluster" (Everitt et al. 2011, 92). Everitt et al. (2011, 92.) point out that a second type of distortion exists, called 'space-dilation' which takes place "where the process of fusing clusters tends to draw clusters together" (Ibid.). Figure 29 illustrates such a *space-dilation effect*, of which we will also be wary.

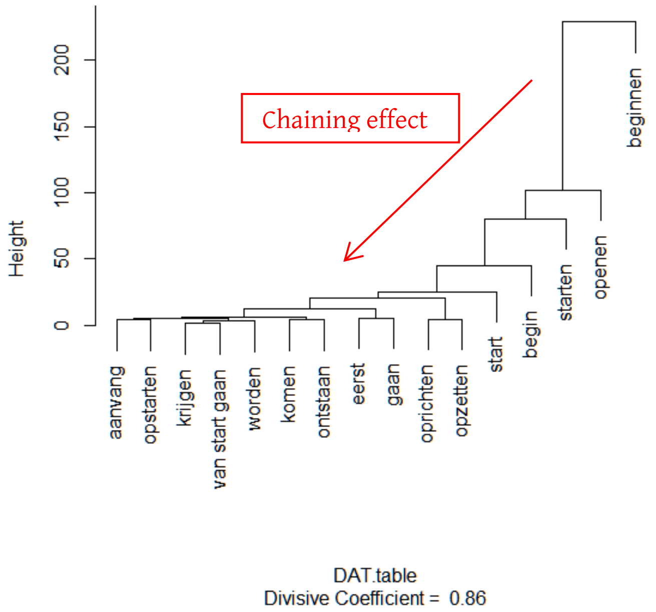


Figure 28 Divise clustering of the field of TransDutch$_{ENG}$, displaying chaining effect

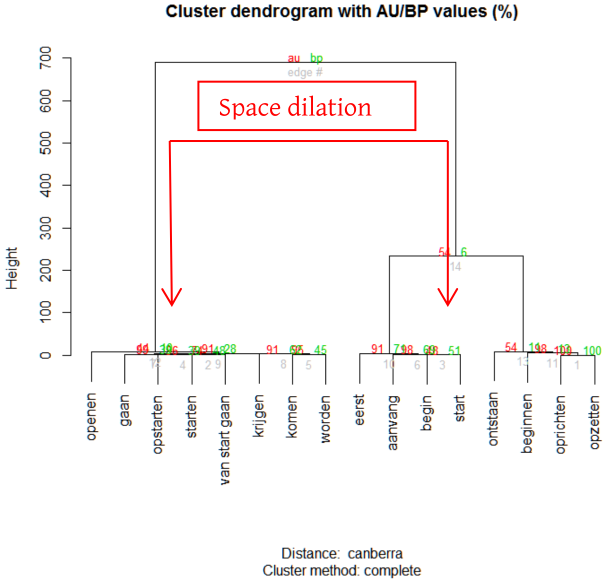


Figure 29 Agglomerative clustering of the field of SourceDutch, displaying space-dilation

As a consequence, we will continue our exploration of the data with *hierarchical agglomerative clustering* (HAC). In addition, we decide to carry out the HAC on the resultant coordinates of the CA. We thereby follow Lebart & Mirkin (1993, 335) who suggest “to complement it [a CA] with a classification”, as this “can supply elements of information that could have been hidden by the projection onto a low dimensional subspace” (see also Ciampi et al. 2005, 28). A HAC performed on the output of a CA has obvious advantages as CA involves dimension reduction: noisy dimensions are omitted and only informative dimensions are retained. By selecting only the informative dimensions of the CA as input for the HAC, our analysis is likely to be better interpretable than a HAC on raw data. In other words, this procedure ‘combines the best of two worlds’: CA allows us to detect informative dimensions of variation to the detriment of noise, and HAC enables us to discern meaningful structure in our data cloud.

Since we will use the output of the CA as input for a HAC, we use the `fast_sca()` function of the `svs`-package to obtain the coordinates (the coordinates can also be obtained by applying the `ca()` function in `R`). The `svs`-function `fast_sca()` is especially designed to further use the resultant coordinates of a CA as the input for an additional analysis.

3.6.2.1 (Dis)similarity measure

Clustering algorithms depend crucially on similarity which is understood as “its everyday meaning of how similar entities are” (Divjak & Fieller 2014, 411). For numerical variables similarities are often converted into dissimilarities (or distance). This can be done by subtracting the measure of similarity from 1. In this way, 0 indicates minimum dissimilarity and 1 maximum dissimilarity (Divjak & Fieller 2014, 415-416).

There is a wide variety of distance measures, and we will not to compare *all* of them[39]. We will limit our comparison to two measures which are customarily used in linguistics: the Euclidean distance and the Canberra distance, the latter is known to handle sparse data and zero-occurrences best (Divjak 2010, 132). Based on the outcome of the comparison (which will be presented in section 3.6.2.4) we choose to use Euclidean as the distance measure for our analyses.

3.6.2.2 Clustering Algorithm

Next to an appropriate distance measure, a clustering algorithm also depends on a so-called 'amalgamation rule'. This determines "which clusters are merged in each step in bottom-up clustering" (Manning & Schütze 1999, 503). In fact, the amalgamation rule is the defining feature of the various agglomerative cluster algorithms as it specifies in which way the proximity between two clusters will be computed; "the definition of cluster proximity that differentiates the various agglomerative hierarchical techniques" (Tan et al. 2006, 517). The most important cluster algorithms are the following:

Single-link clustering (also called nearest neighbor or single linkage algorithm) considers the similarity between two clusters as "the similarity of the two closest objects in the clusters" (Manning & Schütze 1999, 503). This algorithm is known to produce locally coherent clusters, but with a bad global quality (Ibid., 503); the clusters moreover tend to show a *chaining effect* (Ibid., 504).

Complete-link clustering (also called furthest neighbor or complete linkage algorithm) "focuses on global cluster quality [...]. The similarity of two clusters is the similarity of their two most dissimilar members" (Manning & Schütze 1999, 505). This algorithm is known to avoid *chaining effect*, which is preferable in NLP applications (Ibid., 506).

Group-average agglomerative clustering (or average linkage) is a 'compromise' between the previous two algorithms, which uses the average similarity as a criterion to merge items into clusters (Manning & Schütze 1999, 507). It can be considered as an alternative to complete-link clustering and it is also known to avoid *chaining effect*.

Ward's Minimum Variance Method is a somewhat different clustering algorithm as it "allows two clusters to merge if the increase in sum of squared distances[40] of the

[39] the `dist()` function in `R` allows the user to choose between `"euclidean"`, `"maximum"`, `"manhattan"`, `"canberra"`, `"binary"` or `"minkowski"`, the `Dist()` function allows to furthermore use `"pearson"`, `"abspearson"`, `"correlation"`, `"abscorrelation"`, `"spearman"` or `"kendall"`.

[40] The sum of squares is a measure of variation, calculated by summing the squares of the differences from the mean.

members of the new cluster from their mean is smaller than for any other possible merger between two clusters. Use of squared distances penalizes spread out clusters and so results in compact clusters without being as restrictive as complete linkage" (Divjak & Fieller 2014, 426). Because of its tendency to find spherical clusters, Ward's Method is "a frequently recommended strategy that yields small clusters" (Divjak 2010, 133).

The above mentioned algorithms can themselves be grouped according to the different 'views' on clusters they reflect. Depending on the goals one defines, different types of clusters can be found useful. Tan et al. (2006, 493-495) distinguish five types of cluster solutions: well-separated clusters (each object in a cluster is closer or more similar to every other object in the cluster than to any object not in the cluster), prototype-based clusters (each object in the cluster is closer or more similar to the prototype that defines the cluster than to the prototype of any other cluster), graph-based clusters (nodes are seen as objects; the links represent connections among the objects), density-based clusters (a cluster is seen as a dense region of objects surrounded by a region of low density) and shared-property clusters (also called conceptual clusters, where a cluster is a set of objects that share some property). Single linkage, complete linkage and average linkage algorithms suit a graph-based view of clusters; Ward's Method, on the other hand, is the more natural choice when one adheres a prototype-based view on clusters, since it "assumes that a cluster is represented by its *centroid* [...]" (Tan et al. 2006, 517).

Which cluster algorithm is the 'right' one for our purpose, is not a trivial question, as different algorithms yield different dendrograms. Divjak (2010, 132), following Speece (1994/1995, 35) emphasizes to choose the algorithms whose " "side-effects" of the mathematical properties [...] fit the phenomenon under investigation, and, consequently, yield easily interpretable results".

We will discard the single-linkage method because of its tendency to produce *chaining effect*. For the other cluster algorithms, however, it is not so clear which method is preferable. From the previous descriptions, Ward's Method seems to suit our needs best: it can yield small clusters – as a "side-effect of its mathematical properties" – and it reflects a prototype-based view on clusters. The choice of Ward's Minimum Variance method is also what results from the comparison with the complete and average linkage methods in section 3.6.2.4. Hierarchical Agglomerative Clustering is carried out on the output of the CA with the function `pvclust()`from the package `pvclust` which relies on the function `hclust()` (our choice of `pvclust` will be substantiated in the next section).

3.6.2.3 Number of clusters

An important issue of HAC concerns the choice of the number of clusters, i.e., the 'optimal cluster solution'. This is obtained by 'cutting' the tree at a particular height into *n* clusters. The height of the tree cut must be chosen carefully, as the resulting clusters will be considered as meaningful and informative in the subsequent interpretation. The problem is, however, that there does not exist one straightforward procedure to determine the 'best cut'. As a rule of thumb, several scholars suggest that looking at the length of the vertical lines in the dendrogram is indicative for the 'optimal cluster solution'. Gries mentions that "large vertical lines indicate more autonomous subclusters"(2013, 338). Similarly, Divjak & Fieller (2014, 430) propose to "look at the height bar and choose a place where the cluster structure remains stable for a long distance". Finally, Everitt et al. (2011, 95) assert that "large changes in fusion levels are taken to indicate the best cut". Divjak & Fieller admit that such suggestions are not exactly what we would call "frivolous" (2014, 430). To somewhat remedy this, they mention three criteria which can help to make a decision on the cut height. A 'good' cut height should give (i) enough clusters in the solution for it to be meaningful (i.e. an acceptable size); (ii) an immediately intuited meaning for each/most of the clusters and (iii) criterion validity (the expected level of association between rows and columns should be acceptably reflected). Divjak & Fieller (2014, 432-433) furthermore propose two ways to investigate the robustness of a cluster solution (i) the computation of the *average silhouette width* and the use of bootstrap validation.

We decide to determine the optimal cluster solution by means of a bootstrap validation technique (we will use *average silhouette width* as a cluster validation technique, as explained further on in this section). Bootstrapping entails that the data are resampled (with replacement) a high number of times (i.e. usually 3000) in order to see how many times the same points are clustered together again. On the basis of these (3000) replications a p-value is computed for each node of the dendogram (i.e. the place where two branches join). As a consequence, the bootstrap p-values represent a measure of quality for each node. This bootstrap validation will be done with the `R` package `pvclust` (Suzuki & Shimodaira 2006). As a matter of fact, the `pvclust` package provides both an "approximately unbiased p-value" and a "bootstrap probability" (the use of the former is recommended by Suzuki & Shimodaira). In addition, the package has the function `pvrect` which can be used to cut the dendrogram at the nodes above a certain confidence level, e.g. 95%. This has a clear advantage over tree cuts at a fixed height. Fixed-height cuts are common in HAC but not indispensable. Everitt et al. warn that fixed-height cut methods require pre-established cut heights and minimum cluster size which can possibly be influenced by *a priori* expectations (2011, 95).

If possible, we will always *prefer* to cut the tree at the *highest* significant node attaining a confidence level of 95% (as this is in fact the default of how `pvrect` works). However, this procedure runs the risk of excluding many-cluster-solutions: e.g. if the two highest nodes in a tree are significant, `pvrect` would choose a two-cluster-solution. Such solutions with very few clusters might come across as less interpretable.

As a consequence, we propose a compromise of cutting a dendrogram at a confidence level and cutting it at a fixed height: the cutoff point will be chosen so that for each cluster in the solution, the highest node within each cluster is significant (the Approximately Unbiased p^{-}value should be ≥ 0.95) (an exception is made for singleton clusters). In this way our validated cluster solution meets the first two criteria mentioned by Divjak & Fieller for good cut height (acceptable cluster size and meaningful clusters).

Validation of the number of clusters

In the first part of this section we have proposed to use bootstrap p-values in order to determine the number of clusters. We now complement that procedure with two validation techniques for testing the validity of a cluster solution. The first validation consists in the computation of the *average silhouette widths* proposed by Kaufman & Rousseeuw (1990), the second one is a (non-hierarchical) K-means clustering.

Kaufman and Rousseeuw (1990) propose to calculate the *silhouette width* for each object in a cluster solution and summarize this information in a *silhouette plot*. For each object *i*, one can "compare *i*'s separation from its cluster against the heterogeneity of the cluster" (Everitt et al. 2011, 128). The *silhouette width* has a value situated between -1 and 1. Values close to 1 imply that "the heterogeneity of object *i*'s cluster is much smaller than its separation and object *i* is taken as 'well classified'" (Everitt et al., Ibid.); values close to -1 imply misclassification and values around 0 suggest that the classification is unclear (Ibid.). Finally, the *average silhouette width* – the average of all silhouette widths of a set of data – can be used to validate the chosen cluster solution. Kaufman and Rousseeuw point out that an *average silhouette width* above 0,5 indicates a good classification, whereas values beneath 0.2 betray an unclear classification. In addition, Everitt et al. (2011, 129) suggest using the *average silhouette widths* as an instrument for optimizing the number of clusters. We will do the same whenever the average silhouette width of a cluster solution is below 0.5. In such a case, we will compare the average silhouette widths of the cluster solution with K clusters to the average silhouette widths of both the solutions with K-1 and K+1 clusters. The average silhouette width can be calculated using the `pam()` function of the `cluster`-package.

Although K-means clustering can be run as a separate clustering procedure, we will use it as a validation of the HAC. More specifically, we will compute the centers of the clusters from the HAC and feed those into a K-means clustering. If the partitioning of the lexemes into clusters remains (largely) the same in the K-means clustering, then we can regard this as a validation of the results in the HAC. After calculation of the cluster *centroids* using `centers_ca()` function of `svs`, K-means clustering can be carried out using the `kmeans()` function. In contrast to HAC, which does not need a pre-determined number of clusters, other *non*-hierarchical clustering methods such as K-means clustering require a pre-specified number of clusters. More specifically, K-means "defines the clusters by the center of mass of their members" (Manning & Schütze 1999, 515), i.e. it takes K points as the centers of the clusters. For the initialization of the K-means algorithm, K points can be randomly chosen from the data to serve as seeds, although predetermined centers can also be supplied (Ibid.). The algorithm then consists in iteratively assigning each data point to the cluster to the center of which it is closest (Ibid.) and subsequently recomputing the centers on the basis of the assignments (Manning & Schütze 1999, 515-516). This iterative procedure is carried out until convergence, i.e. until there are no further reassignments.

3.6.2.4 Comparison of the chosen procedure with alternative procedures via an assessment of the overall strength of the clustering structure

In the previous sections, we outlined how cluster analysis depends on the choice of distance measures and amalgamation rules. We indicated which distance measure(s) and amalgamation rule(s) were most likely to yield interpretable results for our data. In this section, we will assess various combinations of distance measures (Euclidean and Canberra) and amalgamation rules (Average, Complete, Ward's) on different spatial maps in order to see which combination works best.

In section 3.6.2, we substantiated our choice to carry out a HAC on the output of a CA. Next to this procedure, it is also possible to carry out a HAC directly on the raw data or to compute the distances for the HAC on the output of a Latent Semantic Analysis. LSA is typically considered as a Vector Space Model since "the values of the elements are derived from event frequencies" (Turney & Pantel 2010, 144) and it is also generally associated with distributional approaches to meaning (Ibid., 141). Conceptually, LSA works as follows:

> LSA projects document frequency vectors into a low dimensional space calculated using the frequencies of word occurrence in each document. The relative distances between these points are interpreted as distances between the topics of the documents (Leopold 2007, 123).

LSA can, by virtue of its symmetry, also be applied to word similarity (Leopold 2007, 123) and consequently also to translational similarity. In our case, the algorithm of LSA (which is usually applied to a document-term matrix) is now applied to our matrix, i.e. a source language – target language matrix.

In the subsequent comparison, we will include these two possibilities (HAC on the raw data and HAC on the output of a LSA). The various combinations of distance measures (Euclidean and Canberra), amalgamation rules (Complete, Average and Ward's) and spatial maps (raw data, output of CA, output of LSA) are summarized in Table 13. Because of the high number of combinatorial possibilities – 18 in total – we only apply the comparison to SourceDutch. We selected three validation criteria which have in common their ability to assess the overall strength of the clustering structure.

We calculate the *agglomerative coefficient* for each combination. This is a standard measure to describe the strength of a clustering structure.

> The agglomerative coefficient (AC) [is] a measure of the clustering structure of the data set that can range from 0 to 1. An AC close to 1 indicates that a very clear structuring has been found whereas an AC close to 0 indicates that the algorithm has not found a natural structure. This measure is sensitive to sample size, i.e. the value grows with the number of observations (Divjak & Fieller 2014, 426).

Since we are using the same data set for each dendrogram in this comparison, the agglomerative coefficients will be comparable. We consider an agglomerative coefficient higher than 0.80 as satisfactory.

Table 13 Combinatory possibilities of the selected distance measures, clustering algorithms and 'spatial maps'

	Procedural combination	Agglomerative coefficient	Chaining effect[41]	p-values[42]
1	Euclidean, Average	0,72	YES	10
2	Euclidean, Average, on CA	0,74	YES	10
3	Euclidean, Average, on LSA	0,61	YES	8
4	Euclidean, Complete	0,73	YES	10
5	Euclidean, Complete, on CA	0,76	YES	9
6	Euclidean, Complete, on LSA	0,65	high	4
7	Euclidean, Ward's	0,78	YES	9
8	Euclidean, Ward's, on CA	0,89	NO	9
9	Euclidean, Ward's, on LSA	0,72	NO	4
10	Canberra, Average	0,22	high	2
11	Canberra, Average, on CA	0,95	low (+ high space dilation)	6
12	Canberra, Average, on LSA	0,82	NO	6
13	Canberra, Complete	0,27	NO	1
14	Canberra, Complete, on CA	0,99	low (+ high space dilation)	7
15	Canberra, Complete, on LSA	0,99	low (+ high space dilation)	9
16	Canberra, Ward's	0,43	NO	2
17	Canberra, Ward's, on CA	0,99	low (+ high space dilation)	5
18	Canberra, Ward's, on LSA	0,96	NO	3

[41] 'High' means chaining occurs only in the higher nodes, 'low' means chaining occurs only in the lower nodes.

[42] Number of significant p-values (≥0.95) on a total of 14 nodes.

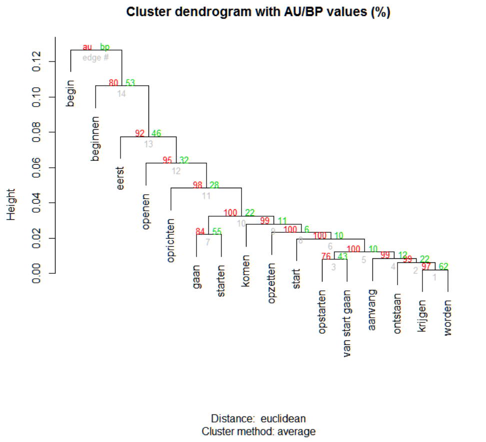


Figure 30 Euclidean, Average (1)

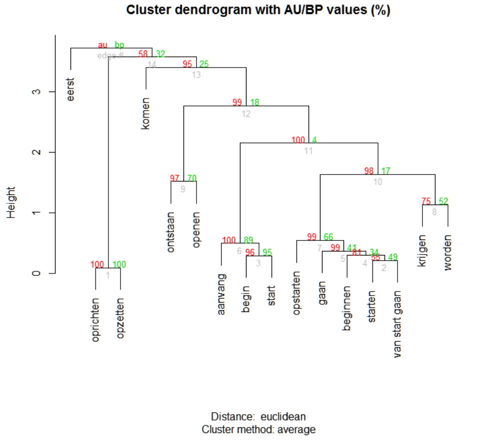


Figure 31 Euclidean, Average, on CA (2)

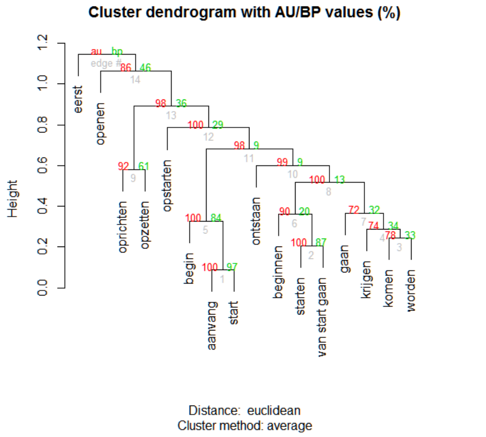


Figure 32 Euclidean, Average, on LSA (3)

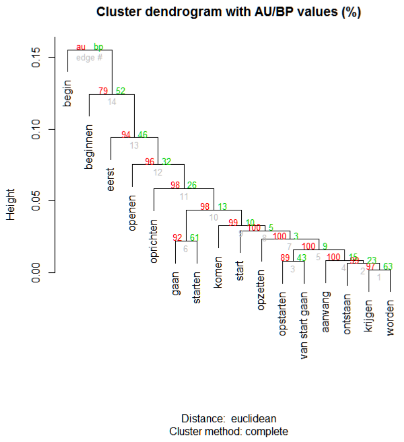


Figure 33 Euclidean, Complete (4)

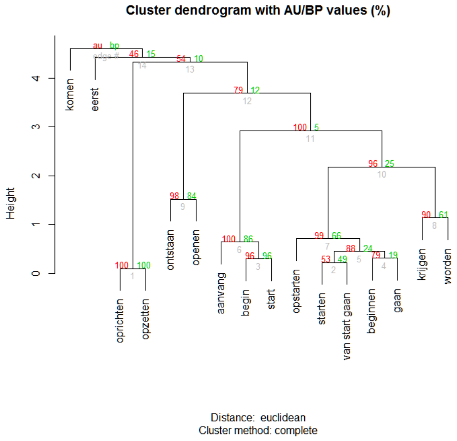


Figure 34 Euclidean, Complete, on CA (5)

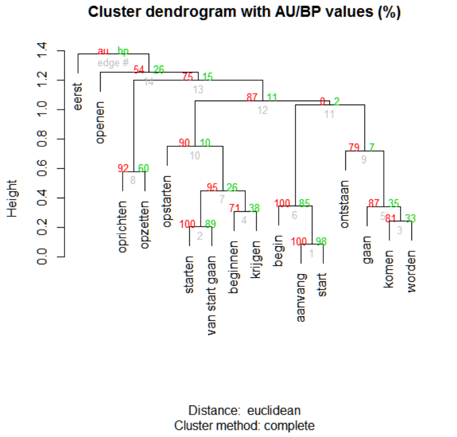


Figure 35 Euclidean, Complete, on LSA (6)

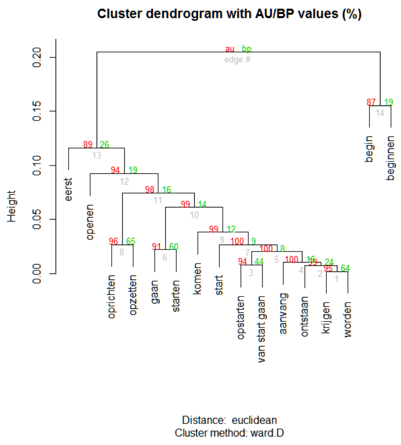


Figure 36 Euclidean, Ward's (7)

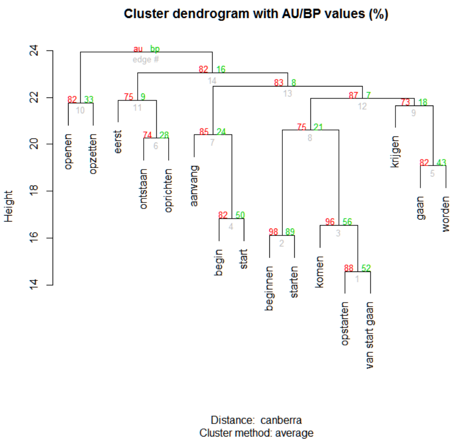


Figure 39 Canberra, Average (10)

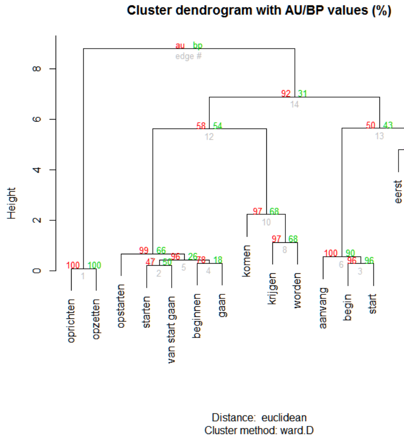


Figure 37 Euclidean, Ward's, on CA (8)

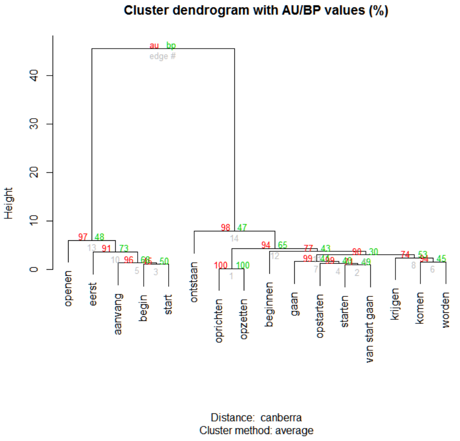


Figure 40 Canberra, Average, on CA (11)

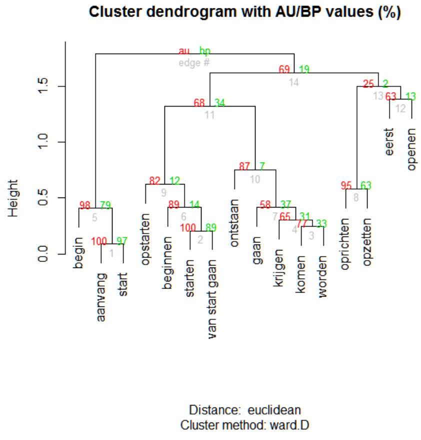


Figure 38 Euclidean, Ward's, on LSA (9)

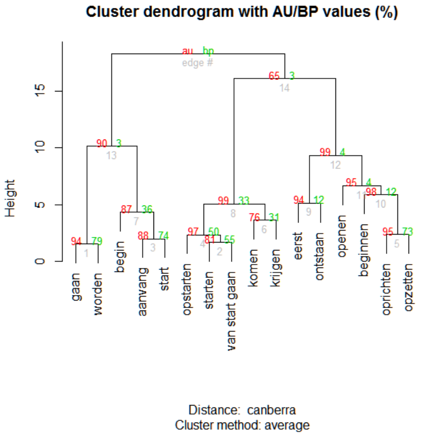


Figure 41 Canberra, Average, on LSA (12)

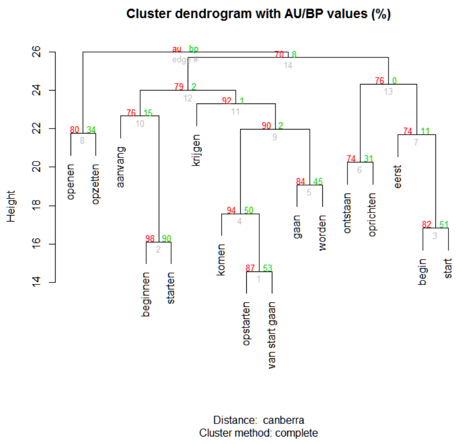


Figure 42 Canberra, Complete (13)

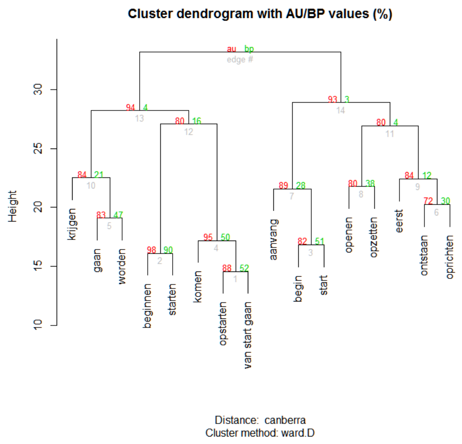


Figure 45 Canberra, Ward’s (16)

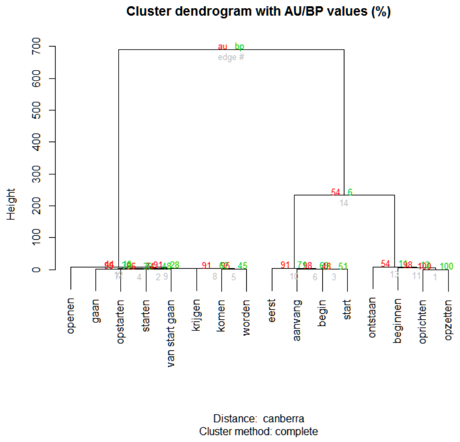


Figure 43 Canberra, Complete, on CA (14)

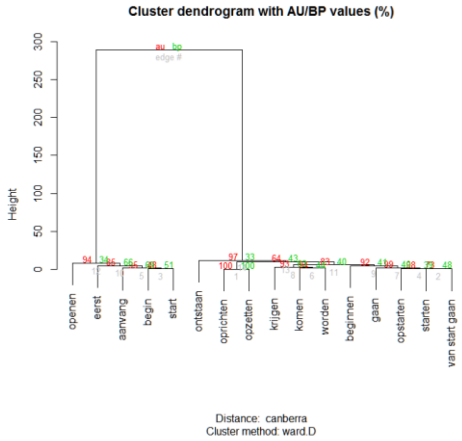


Figure 46 Canberra, Ward’s, on CA (17)

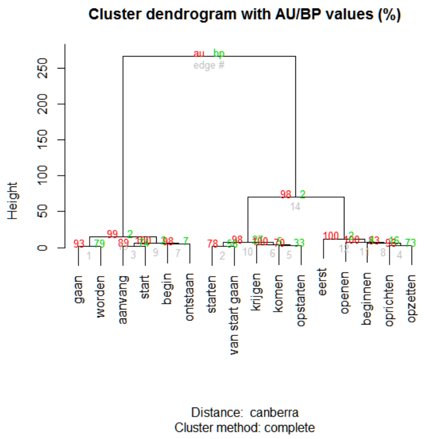


Figure 44 Canberra, Complete, on LSA (15)

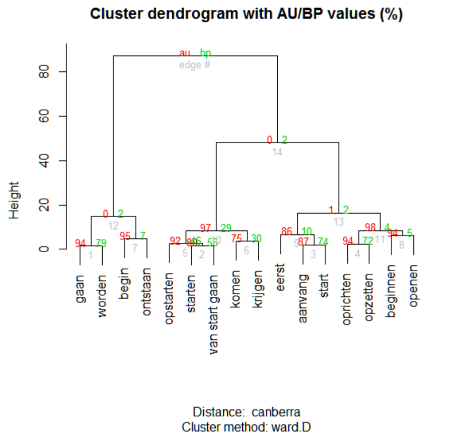


Figure 47 Canberra, Ward’s, on LSA (18)

From Table 13 (and the accompanying Figures 30 to 47[43]), we can read off that combinations 8, 11, 12, 14, 15, 17 and 18 have an agglomerative coefficient higher than 0,80. It is noteworthy that only one combination with Euclidean distance reaches a satisfactory agglomerative coefficient. In addition, we see that for the combinations with Canberra distance, none of the analyses carried out on the raw data display a satisfactory agglomerative coefficient.

We reinforce the assessment on the basis of the agglomerative coefficient by adding two more validation criteria which equally inform us about the cluster structure: *chaining effect* and p-values.

In section 3.6.2, we explained that, for our study, a *chaining effect* in the cluster structure is disfavored to a sphere-like structure. Hence, the appearance of a *chaining effect* (as well as of a *space-dilation effect*) will be considered negative. Because a *chaining effect* can only be determined on the basis of visual inspection, we introduced four levels of chaining. In Table 13, 'no' means that no *chaining effect* was observed, 'yes' that a clear *chaining effect* was observed, 'high' means that chaining occurs only in the higher nodes and 'low' means that chaining only occurs in the lower nodes. Only those results where a clear *chaining effect* is observed ('yes'), will be considered negative, no chaining ('no') will be considered as the most positive outcome.

We see that six out of nine combinations with Euclidean distance show a clear *chaining effect* (combinations 1, 2, 3, 4, 5 and 7). Combination 6 displays chaining on the higher edges of the dendrogram. Only combinations 8 and 9 (using Ward's Minimum Variance Method) do not suffer from chaining. As for the combinations with Canberra distance, we see that none of them displays clear chaining, although combinations 11, 14, 15 and 17 show space-dilation effects on the higher edges as well as chaining-effects on the lower edges. Combination 10 only shows some chaining on the higher edges. Combinations 12, 13, 16 and 18 show no effect of chaining nor space-dilation at all. If we focus on the clustering algorithm, we see that chaining and space-dilation effects are not limited to the complete linkage algorithm but seem to appear irrespective of the clustering algorithm.

Finally, we also use the p-values (which we introduced in section 3.6.2.3 to determine the cluster solution) to assess the overall strength of the clustering structure. We will do this by counting the number of significant nodes (i.e. with a p-value of 0.95 or higher) in the dendrogram. Each of the dendrograms presented in the comparison counts 14 nodes. We will consider ≥ 7 significant nodes as an indication of a strong overall clustering structure. For the combinations with Euclidean distance, we see that all but

[43] For each Figure, the number between brackets refers to the number of the combination in Table 13 it represents. We will use these numbers to refer to the different combinations (not the Figure numbers).

two combinations display a high number of significant p-values (only combinations 6 and 9, carried out on the output of a LSA have less than 7 significant nodes). If we look at the combinations with Canberra distance, we see that only two out of nine combinations have 7 or more significant p-values: combinations 14 and 15, both carried out with the complete linkage algorithm.

On the basis of the obtained values for each of the criteria in the comparison, we can conclude that combinations 8 (Euclidean, Wards, on CA), 14 (Canberra, Complete, on CA) and 15 (Canberra, Complete, on LSA) are most likely to yield interpretable results for our data. Our preference goes to combination 8, because no chaining was observed at all (in combinations 14 and 15 we observed space-dilation in the high nodes and chaining in the low nodes). In addition, this is the only combination with Ward's Method, which is the more natural choice when one adheres a prototype-based view on clusters (as we explained in section 3.6.2.2).

On a more general level, we can conclude that, when Euclidean distance is used, we are more likely to face *chaining effect*, relatively high agglomerative coefficients (although lower than for Canberra) and a high number of significant p-values. Combining Euclidean distance with Ward's Method seems to avoid *chaining effects.* Canberra distance, on the other hand, avoids *chaining effect*, renders high agglomerative coefficients (except on raw data) but renders a low number of significant p-values. From the point of view of the clustering algorithms, it is noteworthy that combinations with the complete linkage algorithm usually display a high amount of significant p-values and that combinations with Ward's Method are usually best at avoiding *chaining effect* (only combination 7 with Ward's displays clear chaining). When we take the different spatial maps as point of departure, we see that analyses on the raw data render low agglomerative coefficients and that analyses on the CA are prone to chaining.

3.6.3 Measuring levelling out via prototypicality effects

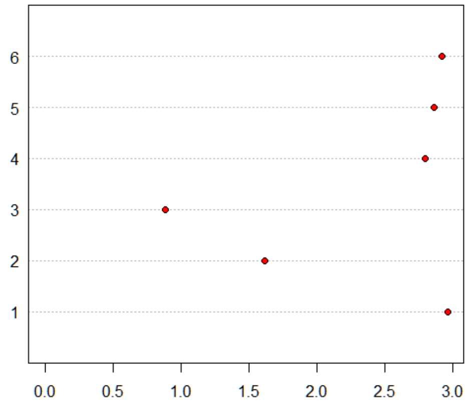
In this section, we want to explore a number of additional statistical techniques to further analyze the structure of the semantic maps yielded on the basis of the HAC. The clusters in a cluster analysis are all on an equal par, i.e. they simply represent a partitioning of the lexemes. However, since we are interested in prototypicality effects (as a proxy for semasiological and onomasiological *levelling out*), we would also like to determine whether certain clusters are more central in the semantic space while others are more peripheral. The measurement of these prototypicality effects will be done on the basis of so-called *centroids* (and to a lesser extent *medoids*), which are calculated on the basis of the coordinates of the CA. Distance to *centroids* will be used to assess the prototype-based organization of clusters within a dendrogram (3.6.3.1), and hence, to investigate semasiological *levelling out*. *Centroids* and *medoids* will also be used to assess the prototype-based organization of lexemes within a cluster (3.6.3.2) to investigate onomasiological *levelling out*. Section 3.6.2.3 will further explore how these two measures may represent different views on prototypes. In addition, each cluster in a dendrogram will also receive a meta-label in an attempt to capture the specific meaning distinction of the cluster (3.6.3.4).

3.6.3.1 Organization of clusters within each dendrogram

We propose to explore the prototype-based organization of the clusters within each dendrogram by assessing the distance of each cluster's center (its *centroid*) to the zero-point of the semantic space.

Centroids correspond to the average of all points in the cluster (Tan et al. 2006, 494). They can be calculated on the resulting coordinates of the CA (recall that the output of the CA will be used as input for the HAC). The *zero-point* or *origin* of a semantic space corresponds to the weighted mean of the columns and of the rows (they are superposed and calibrated on the *zero-point*). If a data point is situated close to the *origin*, this implies that its weighted mean is close to the overall weighted mean. The data point can hence be considered as 'central' in the spatial map, and its profile will be rather resembling to other, equally central points in the spatial map. If we accept Lakoffs (1987) idea that lexical categories and polysemy networks are structured with respect to their prototypical meanings (Tyler & Evans 2003), and if we furthermore accept Dyvik's basic idea that "semantically closely related words ought to have strongly overlapping sets of translations" from which it follows that strongly overlapping sets of translation ought to reveal semantic relatedness; then this leads us to believe that the central sphere of a spatial map – close to the *zero-point* or *origin* – can be considered as the prototypical center. As a consequence, the data points (be it *centroids* or lexemes) which find themselves in or close to this central sphere can then be considered as prototypical points in the semantic space. The distances of the clusters' *centroids* to the zero-point (the prototypical center) of the semantic space they belong to can then inform us about the more prototypical or more peripheral position of each cluster (meaning distinction) in the semantic space (the semantic field it belongs to).

We calculate the coordinates of the cluster center (the *centroid*) on the output of the CA (i.e. the coordinates of the CA) with the built-in function `centers_ca()` from the `svs`-package. We then compute the Euclidean distance from each *centroid* to the zero-point of the semantic space with the helper function `dist_wrt()` from `svs`. Finally, we can visualize the distances of the *centroids* to the origin of the semantic space with a dot chart. The example in Figure 48 shows the distance of each of the clusters in the HAC visualization for SourceDutch to the origin of the semantic space. On the basis of this visualization, we can see which clusters are situated closer to the origin of the semantic space and which ones are more peripheral. Since we consider the zero-point of the semantic space as the prototypical center, we consider clusters that are closer to the zero-point of the semantic space as more prototypical and clusters further away from the zero-point as more peripheral.



Cluster 6	eerst
Cluster 5	krijgen, komen, worden
Cluster 4	ontstaan, openen
Cluster 3	Starten, van start gaan, opstarten, beginnen, gaan
Cluster 2	aanvang, begin, start
Cluster 1	opzetten, oprichten

Figure 48 Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of SourceDutch

3.6.3.2 Organization of the lexemes within each cluster

The prototype-based organization of the different items (lexemes) within each cluster can equally be assessed with *centroids* by measuring the distance of each lexeme to the *centroid* of the cluster it belongs to. The Euclidean distance from the lexemes to each of the cluster *centroids* can be calculated with the function `dist_wrt_centers()` from svs and visualized in a dot chart (an example can be found in Figure 49). The distance of the lexemes to the *centroid* (the average of all points in the cluster) of the cluster they belong to can be used to explore which lexical items are more prototypical expressions of the particular meaning distinction (indicated by the cluster) and which ones are more peripheral. For the example in Figure 49, for instance, we see that *starten* and *beginnen* are the lexemes situated closest to the *centroid* of the cluster they belong to, implying that they are closest the abstract prototype contained in the *centroid* (see section 3.6.3.3).

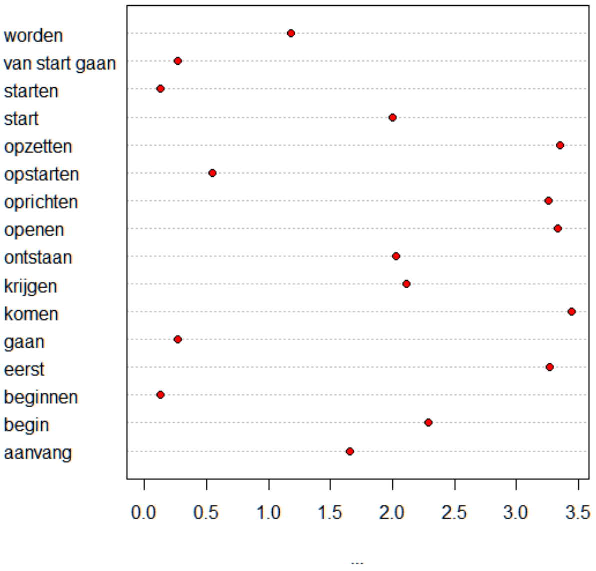


Figure 49 Dot chart representing the distance of the lexemes to the centroid of cluster n°4 for SourceDutch

This analysis will also enable us to determine the stability of the cluster membership of each lexeme. HAC is categorized as *hard clustering*, which means that each object in the analysis can be assigned to only one cluster (in contrast to *fuzzy clustering*, which can reveal the degree of membership of an object to a cluster). By looking at the distance of the lexemes to their cluster's *centroid*, we can somewhat 'nuance' the hard clustering. The positions of the lexemes with respect to their *centroid* may show that some lexemes are 'hesitant' between two clusters, and their assignment to a particular cluster is not as straightforward and clear-cut (as *hard*) as the dendrogram structure would have suggested. The *centroid* itself, however, is not a meaningful point[44] since it is the average of all points. Alternatively, it is possible to compute the *medoid* for each cluster, which is the particular point in the cluster with the smallest average distance to all other points

[44] Manning and Schütze (1999, 516) point out that the centroid "is in most cases not identical to any of the objects".

(Divjak 2010, 164). Everitt et al. note that the term *medoid* was coined by Kaufman and Rousseeuw (1990) by analogy with calling the group mean the *centroid*. The *medoid* "can be interpreted as a representative object or *exemplar* of the group" (Everitt et al., 2011, 113) and is necessarily *one* object in the cluster; this object can then be considered as the "prototypical class member" (Manning & Schütze 1999, 516) in a cluster. The medoid can be calculated with the `pam()`-function in `R` ('Partitioning around Medoids').

For each cluster analysis, we will calculate both the *medoid* of each cluster as well as the distance of each lexeme to the *centroid* of the cluster it belongs to. Both measures seem to have their own advantage(s). The distances of each of the lexemes to the *centroid* allow us to better understand the organization of the lexemes in a cluster as a 'continuum' with some lexemes closer to the *centroid* (the most central ones) and others further away from the *centroid* (the most peripheral ones). The *medoid* on the other hand indicates one particular lexeme but is less informative about the structure of the cluster. If the *medoid* happens to be different from the lexeme closest to the *centroid*, this could indicate tension between several prototypical expressions.

3.6.3.3 Centroids and medoids: different views on prototype

Both measures (distance to the *centroid* and *medoid*) can be used to determine which lexical item in each cluster can be considered as the most prototypical expression of that cluster (the particular meaning distinction indicated by the cluster). However, distance to *centroid* and *medoid* could be seen as representing two different views on prototypes.

Descriptions of the prototype-based organization of the lexical items in a cluster which rely on the distance of the items to the *centroid* imply that we see prototype as a "summary representation" (Murphy 2004, 42), meaning that "an entire category is represented by a unified representation" where "[t]he concept is represented as features that are usually found in the category members, but some features are more important than others" (Murphy, Ibid.). Because such a summary representation is (always) abstract, it would *strictu sensu* not be possible to capture the summary representation within only one lexeme of the cluster (since the prototype would be an abstract sum of features). We could, however, consider the *lexeme closest to the centroid* as the one that – in the best way possible – reunites the features usually found in the category members, without considering it as the 'ideal member' (the ideal member would be the *centroid* itself, which does not coincide with any of the cluster's members). Hence, the *lexeme closest to the centroid* can be seen as the best possible representation of the abstract prototype contained in the *centroid*. If we consider the *medoid* of a cluster as the prototype of the cluster it belongs to (the particular meaning distinction), this would imply that we adhere to what Murphy calls the "best example idea" (2004, 42), where "a single prototype could represent a whole category" (Murphy, Ibid.). The *medoid* then indicates the best example as the prototype of the cluster it belongs to.

3.6.3.4 Manual assignment of meta-labels

In the introduction of section 3.6.3, we announced that we want to assign a meta-label to each cluster in the dendrogram so as to name the specific meaning distinction indicated by the cluster. There are several options to arrive at such a label. Firstly, we could decide to select either the lexeme closest to the *centroid* or the *medoid* of each cluster as its meta-label. However, since only 16 lexemes will be making up our dendrograms, we can expect several small clusters (with 3 or fewer members) to appear. Indicating one of the few lexemes in such a small cluster as its meta-label will most likely not have much informative value with respect to the specific meaning distinction of that cluster.

Secondly, we could apply other quantitative techniques to provide us with supplementary information about each cluster. This would, however, require an expansion of the amount and nature of annotated data in our data sets. It is possible, for instance, to carry out a supplementary annotation (e.g. of contextual information) and to add this information to the analysis. One possibility would be to apply a behavioral profiling (Divjak & Gries 2006, 2008, Gries & Divjak 2009) to the resulting data sets (which consists in coding each item occurring in each of the sentences for a number of variables, known as ID tags). While such an analysis would have certainly yielded new insights, such a laborious and time-intensive task could not be carried out within the scope of this study.

A third option is to manually label each cluster in an attempt to capture its specific meaning distinction via a more qualitative analysis of each cluster. For this study, we will opt for such a manual assignment task, which will consist in a thorough inspection of each cluster in a dendrogram. The assigned meta-label will combine information of three types of sources: corpus examples from the DPC containing the lexemes which make up a cluster, attestations in reference works and information from the lexical database Cornetto (Vossen et al. 2008; 2013). Cornetto is a lexical data base for Dutch which consists of two existing semantic resources (Dutch Word Net and Referentiebestand Nederlands). It was created within the same project (STEVIN) as the Dutch Parallel Corpus that we are using in this study (see section 3.2). The semantic properties of words are described in Cornetto by the categories Sentiment (with labels such as ‘positive’ and ‘negative’), Pragmatics (including usage information about domain, chronology, connotation, geography and register), Semantics (with specific values for each part-of-speech) and SenseExamples (information about the combinatoric properties). It must be admitted that the integration of the variety of semantics-related information obtained via Cornetto could also have been done in a quantitatively more robust way, rather than via the qualitative analysis proposed above[45]. However, such an operation would have (again) required an expansion of the amount and nature of annotated data (the resulting data sets of the SMM++ would need supplementary annotation with the semantic information from Cornetto before an analysis using those tags as variables could be carried out). Although such an analysis would definitely enrich the dendrograms and consequently allow for more fine-grained descriptions of the clusters – while simultaneously adding interpretative power – we did not further investigate this option within the purview of this study (because our first concern was to explore as many potentialities as possible of translational data 'alone' for semantic description, without using any additional annotative information in the analysis).

3.6.4 Measuring shining through on the semantic level

In this section, we propose two additional visualization tasks which will allow us to investigate semasiological and onomasiological *shining through.*

The investigation of semasiological *shining through* requires us to compare the meaning distinctions in each dendrogram representing translated language to meaning distinctions present in the source language of the translation. In order to carry out such a comparison, we will visualize the semantic fields of the closest equivalents of *beginnen* in the source languages of TransDutch$_{\mathrm{ENG}}$ and TransDutch$_{\mathrm{FR}}$, viz., SourceEnglish *to begin* and SourceFrench *commencer.* We will compare the different meaning distinctions (clusters) in the fields of *to begin* and *commencer* to the meaning distinctions in translated language (TransDutch$_{\mathrm{ENG}}$ and TransDutch$_{\mathrm{FR}}$) to see whether the specific (prototype-based organization of the) meanings within the semantic fields of SourceEnglish and SourceFrench have possibly influenced the organization of the meanings in translated language (TransDutch$_{\mathrm{ENG}}$ and TransDutch$_{\mathrm{FR}}$). The resulting semantic spaces of inchoativity in French and English are *independent* of TransDutch and correspond to the *second T-images*[46] of *commencer* and of *to begin.* These additional

[45] A quantitatively more robust way of integrating this variety of informative semantics-related labels into the analysis would be to manually tag the resulting data sets of the SMM++ with the semantic information from Cornetto and carry out an analysis using those tags as variables (as an alternative analysis to the clustering on the basis of translations/source language lexemes). Another option would be to add the information of these semantics-related labels as supplementary points to a Correspondence Analysis based on the translational data. Thirdly, one could also envisage to use the previously obtained translational information as an additional tag and carry out a cluster analysis using both the semantics-related labels and the translations as variables.

[46] Note that for *commencer* and *to begin*, only one mirroring can be carried out (i.e. with a single language B–Dutch) since the DPC does not contain the translation directions French-English, English-French.

semantic spaces can serve as useful reference points to tease apart target language (*normalization* effects) and source language influence (*shining* through).

Onomasiological *shining through* can be investigated by visualizing the English and French source language lexemes (which determine the clustering of the Dutch lexemes in TransDutch$_{ENG}$ and TransDutch$_{FR}$ into specific meaning distinctions) together with the Dutch target language lexemes. In this way, we can see whether the specific organization of the lexical items in the clusters – with each cluster representing a particular meaning distinction of *beginnen* – is possibly influenced by a specific underlying source language lexeme. In order to obtain a simultaneous representation of the source and target language lexemes in a single semantic space, we will carry out a Multiple Correspondence Analysis on a Burt table (Greenacre 2006, 2007). Burt tables are generalizations of ordinary frequency tables with row and column categories, in that they cross all categories as rows with all categories as columns. The advantage of a Multiple Correspondence Analysis on a Burt table is that distances can be computed, not only between (Dutch) target lexemes themselves, but also between target lexemes and source lexemes so that both source language lexemes and target language lexemes are represented in a single space. This MCA on a Burt table is subsequently visualized with a HCA, enabling us to visually inspect which Dutch target lexemes are associated with which French or English source lexemes.

3.7 Conclusion

In this chapter, we have pursued two goals. The first one was to establish a translation-driven retrieval method for the selection candidate-lexemes for a semantic field. By means of a first extension of the SMM, we developed such a retrieval method based on the different translational statuses (either source or target language) of data.

The second goal was to arrive at a visualized representation of the retrieved data sets. To this end, we combined Correspondence Analysis with Hierarchical Cluster Analysis. We applied CA in order to construct a low-dimensional semantic space of our data. Subsequently, we applied HAC in order to find more structure in our semantic spaces in the form of clusters of lexemes. We calibrated our technique by the Euclidean

Consequently, the data sets for the *second T-images* are based on a single data set (compared to the *second T-image* data set for SourceDutch, which consists of the combined data of the *second T-image* of beginnen$_{FR}$ and beginnen$_{ENG}$).

distance measure and Ward’s Minimum Variance Method as the amalgamation rule, which will be applied in the subsequent chapters to all data sets retrieved in sections 3.5.1 and 3.5.2.

The comparison of the visualizations representing the semantic fields of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ will allow us to investigate *normalization* effects in translation. In this section, we also proposed a number of additional analyses which will enable us to investigate the universals of *levelling out* and *shining through* on both the semasiological and the onomasiological level.

The visualizations which ensue from the application of the developed method are presented in the next chapter.

Chapter 4
Results

4.1 Introduction

In this chapter, we will first present a number of visualizations (one for SourceDutch, one for TransDutch$_{ENG}$ and one for TransDutch$_{FR}$) that were yielded on the basis of the methodological procedure developed in the previous chapter (see section 3.6.2). For each visualization, we will calculate (i) the distance of each cluster's *centroid* to the zero-point (considered as the prototypical center) of the semantic space it belongs to, (ii) the distances of the lexemes in each cluster to their cluster's *centroid* (considered as the abstract prototype of the cluster) as well as (iii) the *medoid* of each cluster (considered as the best exemplar of the cluster). The distances of the *centroids* to the zero-point of the semantic space (the prototypical center) inform us on the semasiological level about the prototype-based organization of the clusters (the meaning distinctions) in the semantic space (the semantic field of *beginnen*). The distances of the lexemes to the *centroid* of the cluster they belong to give us more information on the onomasiological level about the prototype-based organization of the lexemes within each cluster. The *medoid* (the best exemplar) as well as the lexeme closest to the *centroid* of a cluster (the best representation of the abstract prototype) can be used to determine the most prototypical expression in each cluster.[47] We will provide an in-depth interpretation of each visualization representing a semantic field of *beginnen* / inchoativity. 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 meta-labels that we will assign should be understood as a post-hoc, interpretative tool, applied to enhance our understanding of the rendered dendrograms.

[47] See section 3.6.3 for the rationale behind the use of *centroids* and *medoids* to approach prototypicality.

The obtained visualizations and interpretations will be used as a basis to investigate the universal tendencies of *levelling out*, *shining through* and *normalization* on the semantic level.

Semasiological *levelling out* ("do the meanings expressed by *beginnen* differ in translated language compared to non-translated language?") will be investigated by comparing the prototype-based organization of the clusters in each dendrogram (SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$) to each other. The prototype-based organization of the meaning distinctions (the clusters) is evaluated on the basis of the distance of each cluster's *centroid* to the zero-point of the semantic space it belongs to (considered as the prototypical center). Possible changes in the distances of the clusters' *centroids* to the prototypical center amongst the different varieties, can be used to evaluate the organization of those meanings in translated Dutch compared to non-translated Dutch. If these changes consist in *beginnen* having fewer different meaning differentiations in translated Dutch compared to *beginnen* in non-translated, we can call the phenomenon semasiological *levelling out*.

Onomasiological *levelling out* ("do the lexical expressions used to express the different meaning distinctions of *beginnen* differ in translated language compared to non-translated language?") will be investigated by comparing the prototype-based organization of the lexemes in each cluster and for each field (SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$) to each other. This can be done by evaluating the distance of each lexeme to the *centroid* (considered as the abstract prototype) of the cluster (the meaning distinction) it belongs to. For each visualization, we will determine the distances of all the lexemes to the *centroid* of their cluster. Changes in the distances of the lexemes to their *centroids* can inform us about differences in the prototype-based organization of those lexemes in translated Dutch compared to non-translated Dutch.

Semasiological *shining through* (source language influence on the meaning distinctions in translated language) will be investigated by comparing the meaning distinctions in translated language to those present in the source language of the translation. This can be done by visualizing the semantic fields of the closest equivalents of *beginnen* in the source languages of TransDutch$_{ENG}$ and TransDutch$_{FR}$, viz. SourceEnglish *to begin* and SourceFrench *commencer*. By comparing the meaning distinctions in the fields of *to begin* and *commencer* to those present in translated language (TransDutch$_{ENG}$ and TransDutch$_{FR}$), we can see whether the specific meaning distinctions within the semantic fields of SourceEnglish and SourceFrench have influenced the organization of the meaning distinctions in translated language (TransDutch$_{ENG}$ and TransDutch$_{FR}$).

Onomasiological *shining through* (source language influence on the prototype-based organization of the lexemes within each meaning distinction of *beginnen* in translated language) will be investigated by visualizing the French and English source language lexemes together with the Dutch target language lexemes. In this way, we can see whether the specific organization of the lexical items in the meaning distinctions in the fields of TransDutch$_{ENG}$ and TransDutch$_{FR}$ is influenced by a specific underlying source language lexeme.

Semasiological *normalization* (target language influence on the meaning distinctions in translated language) can be investigated by comparing the meaning distinctions in translated language to those present in non-translated language. This can be done by comparing the meaning distinctions present in the visualizations of SourceDutch to the meaning distinctions in TransDutch$_{ENG}$ and TransDutch$_{FR}$. If a same meaning distinction appears in TransDutch$_{ENG}$ and TransDutch$_{FR}$ and this organization is in addition similar or identical to the organization in SourceDutch, there is a fair chance that the TransDutch fields are 'conforming' to the SourceDutch field, yielding evidence for semasiological *normalization.*

Onomasiological *normalization* (target language influence on the prototype-based organization of the lexemes within each meaning distinction of *beginnen*) can be investigated by comparing the prototype-based organization of the lexemes in each meaning distinction in translated language to those present in non-translated language. This can be done by comparing the prototype-based organization of the lexemes in each meaning distinction in SourceDutch to the organization of the lexemes in each meaning distinction in TransDutch$_{ENG}$ and TransDutch$_{FR.}$ If the same organization of lexemes appears in TransDutch$_{ENG}$ and TransDutch$_{FR}$ and this organization is similar or identical to the organization in SourceDutch, there is a good chance that the TransDutch fields are 'conforming' to the SourceDutch field, yielding evidence for onomasiological *normalization.*

Our statements about semasiological change will be based on the outcome of a statistical analysis and an interpretation of clusters as meaning distinctions. Conclusions about onomasiological change will be based on measurements of minimal (and hence subtle) differences in distances to an abstract prototype contained in the *centroid*. It should be clear that our attempt to present a post-hoc interpretation of the quantitative and statistical information in terms of semantic change needs to be seen as a first exploration of the field of inchoativity and by no means an endpoint.

The outline of this chapter is as follows. In sections 4.2, 4.3 and 4.4, we will provide a description as well as an interpretation of the visualizations of the semantic field of *beginnen* / inchoativity of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ respectively. Each description will consist of the following elements: (i) the results of the Hierarchical Agglomerative Cluster Analysis (carried out on the output of a Correspondence Analysis), (ii) a description of the prototype-based organization of the clusters in the dendrogram based on the distances of the *centroids* to the zero-point of the semantic space, (iii) a description of the prototype-based organization of the lexemes within each cluster based on the distances of the lexemes in each cluster to their cluster's *centroid* ,

(iv) a description of the *medoid* of each cluster. Finally, (v) an interpretation of each visualization representing a semantic field of *beginnen* / inchoativity will be provided, on the basis of which a meta-label will be determined for each cluster so as to name the specific meaning distinction revealed by that cluster.

In sections 4.5, 4.6 and 4.7 we will present our insights with respect to tendencies of *levelling out*, *shining through* and *normalization* each time on both on the semasiological and on the onomasiological level. The interpretations of the fields of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ described in the previous sections will form the basis here.

4.2 SourceDutch

4.2.1 Results of the Hierarchical Agglomerative Cluster analysis

Following the procedure described in chapter 3, we will carry out a HAC on the output of a CA. We first apply the statistical technique of CA. The scree plots in Figures 50 and 51 show the distribution of the variation over the latent dimensions of the CA. The cumulative scree plot (Figure 51) shows that at least 5 dimensions are needed to represent more than 80% of the variation:

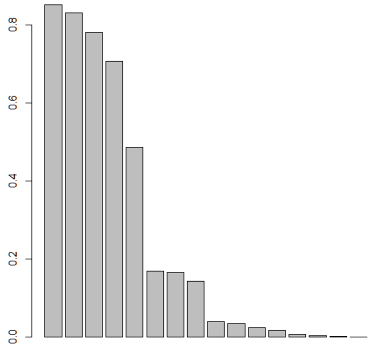


Figure 50 Scree plot for SourceDutch

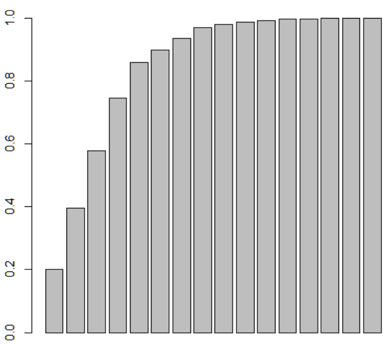


Figure 51 Cumulative scree plot for SourceDutch

On the basis of this scree plot, we reduce the number of dimensions of the CA to 5. This step is important to avoid noisy (less informative) data patterns (see section 3.6). A HAC can now be carried out on the output of the CA. The cut-off point is set at a height of 4 (following the rationale described in section 3.6.2.3)[48], resulting in a cluster solution with 6 clusters: cluster n°1 contains *oprichten* [to establish] and *opzetten* [to set up]; cluster n°2 includes *aanvang* [commencement], *begin* [beginning] and *start* [start]; cluster n°3 comprises *opstarten* [to start up], *starten* [to start], *van start gaan* [to take off], *beginnen* [to begin] and *gaan* [to go]; cluster n°4 holds *ontstaan* [to come into being] and *openen* [to open]; cluster n°5 consists of *komen* [to come], *krijgen* [to get] and *worden* [to become]; cluster n°6 contains *eerst* [firstly]. We consider the result presented in Figure 52 as a possible visualization of a semantic field of *beginnen* / inchoativity in SourceDutch[49].

[48] Note that – had we applied `pvrect()`, which cuts off each cluster at the highest possible node with a significant p-value – the same cluster solution would have been obtained.

[49] For the reasoning behind the assignment of the numerals to the clusters, please refer to section 4.2.2.

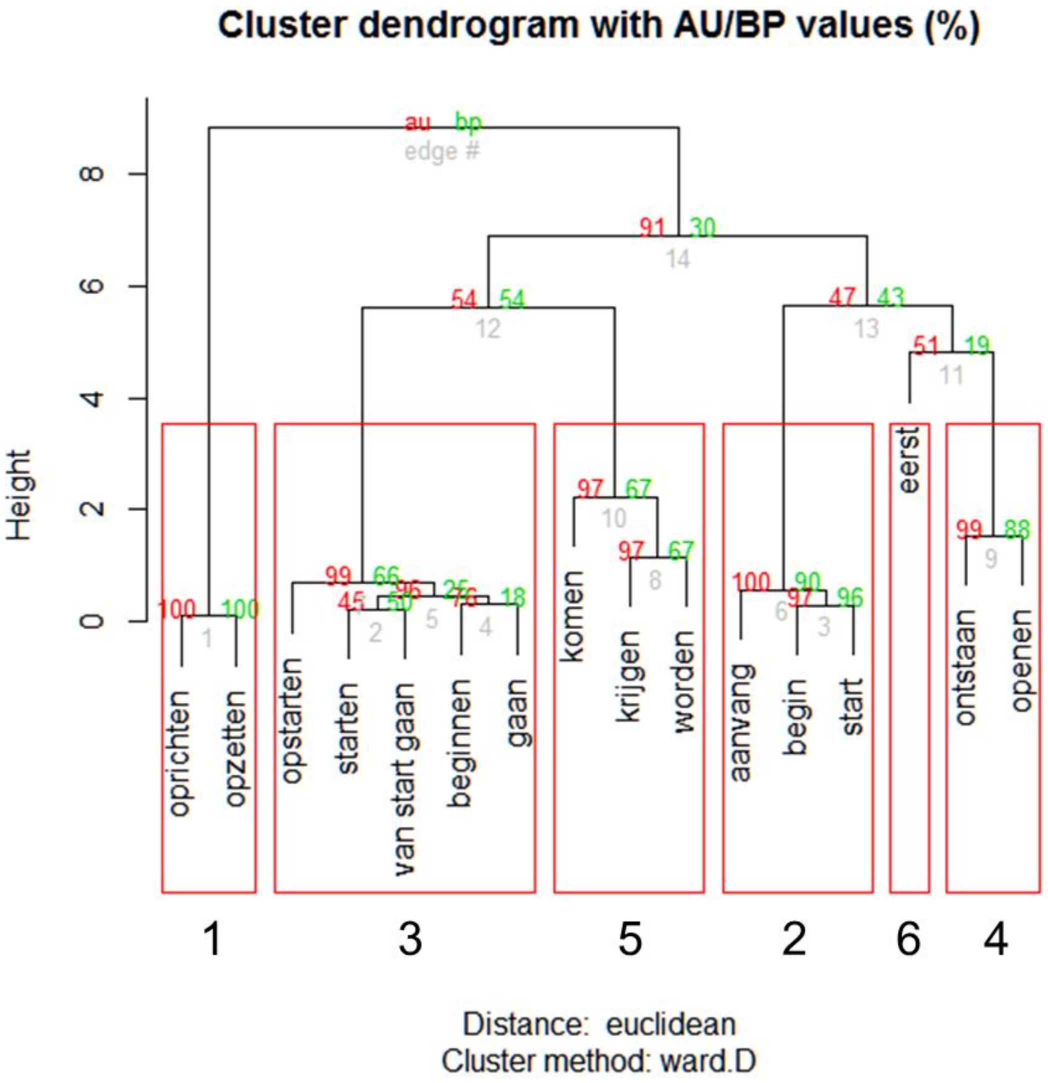


Figure 52 Dendrogram representing a semantic field of *beginnen* / inchoativity for SourceDutch

In order to validate the chosen cluster solution with 6 clusters, we calculate the *average silhouette width*. We obtain an *average silhouette width* of 0.59 for this cluster solution, which is above the 0.50 threshold for good classification determined by Kaufman and Rousseeuw (see section 3.6.2.3).

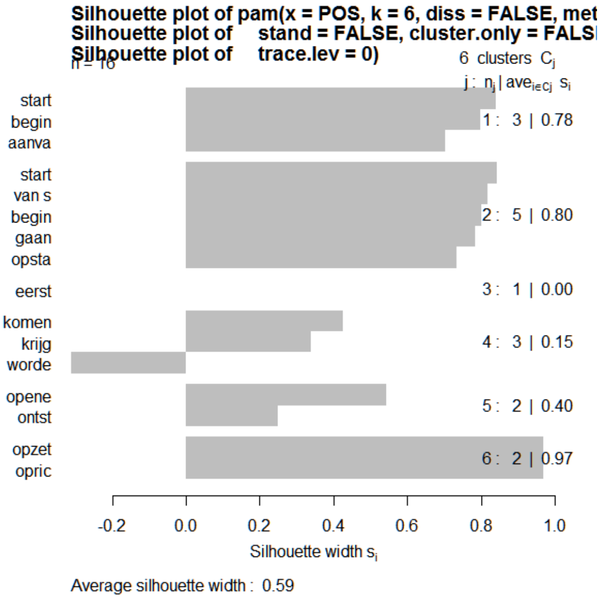


Figure 53 Average silhouette width for cluster solution with 6 clusters for SourceDutch

A K-means clustering is carried out as a second validation technique for the chosen cluster solution. When a cluster solution with 6 clusters is requested, the following K-means clustering is proposed (the numeral beneath each lexeme assigns it to a specific cluster):

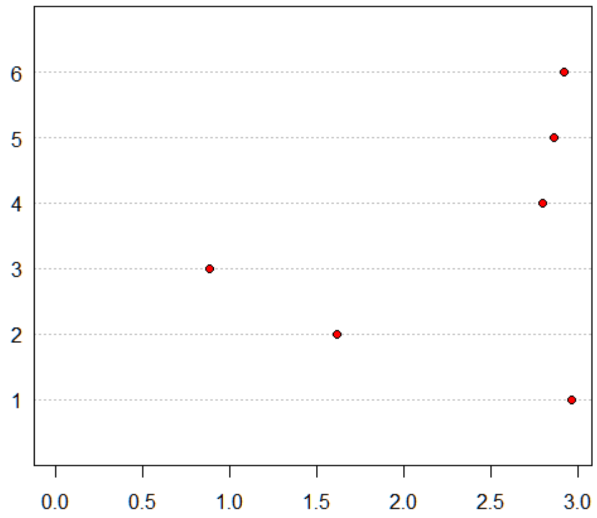
```
Clustering vector:
       aanvang          begin       beginnen          eerst           gaan
             2              2              3              6              3
         komen        krijgen       ontstaan         openen      oprichten
             5              5              4              4              1
     opstarten       opzetten          start        starten van start gaan
             3              1              2              3              3
        worden
             3
```

Note that the only difference with the output of the HAC is that *worden* is assigned to the cluster containing *starten*, *van start gaan*, *opstarten*, *beginnen*, and *gaan*. On the basis of both validation techniques, we can consider our cluster solution for SourceDutch as a good classification. In addition, as a result of the K-means clustering we have found out that the clustering of the polyfunctional verb *worden* seems to be uncertain.

4.2.2 Prototype-based organization of the clusters in the dendrogram (semasiological level)

In order to obtain more information about the prototype-based organization of the clusters (meaning distinctions) within each dendrogram, we determine the distance of the *centroids* of each cluster to the origin or zero-point of the semantic space (the prototypical center). The *centroids* are subsequently mapped onto a dot chart (Figure

54). The cluster closest to the zero-point will be considered as the most central one in the semantic space.



Cluster 6	eerst
Cluster 5	krijgen, komen, worden
Cluster 4	ontstaan, openen
Cluster 3	Starten, van start gaan, opstarten, beginnen, gaan
Cluster 2	aanvang, begin, start
Cluster 1	opzetten, oprichten

Figure 54 Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of SourceDutch

Note that the numerals on the y-axis of the dot chart in Figure 54 are assigned by a previously established list (based on the output of the cluster analysis), necessary to calculate the cluster *centroids* (the order of the assigned numerals is arbitrary). The content of each cluster number is resumed in the table accompanying Figure 54. The dot chart shows us that cluster n°3, containing *starten*, *van start gaan*, *opstarten*, *beginnen* and *gaan* is the most central cluster in the analysis, rather closely followed by cluster n°2 comprising *aanvang*, *begin* and *start*. Then, clusters n°4 (*ontstaan* and *openen*) n°5 (*komen*, *krijgen* and *worden*), n°6 (*eerst*) and n°1 (*oprichten*, *opzetten*) are situated considerably further away but at an almost equal distance of the plot’s origin.

4.2.3 Prototype-based organization of the lexemes within each cluster (onomasiological level)

We now inspect the prototype-based organization of the lexemes within each cluster by measuring the distance of the lexemes of each cluster to the *centroid* of the cluster they belong to. In addition, we calculate the *medoid* of each cluster. As we have seen in section 3.6.3.3, both the lexeme closest to the *centroid* and the *medoid* can be used to determine which lexical item in each cluster can be considered as the most prototypical expression of that cluster (although we regard the two measures as different views on prototypes: the *lexeme closest to the centroid* is considered as the best possible representation of the abstract prototype contained in the *centroid*, the *medoid* indicates the best example as the prototype of the cluster it belongs to).

4.2.3.1 Centroids

Each of the six dot charts (Figures 55 to 60) represents one of the six clusters of SourceDutch. The *centroid* of the represented cluster is taken as the zero-point of the dot chart, so that the lexemes pertaining to this cluster are the closest ones to the zero-point of the dot chart. This permits us to visualize which lexemes are more central, and which ones more peripheral in the cluster. In addition, these visualizations also show the distance of the lexemes of all the other clusters to the represented cluster *centroid*. This is especially interesting for lexemes of which the proposed clustering on the basis of the HAC appeared uncertain (e.g. *worden*).

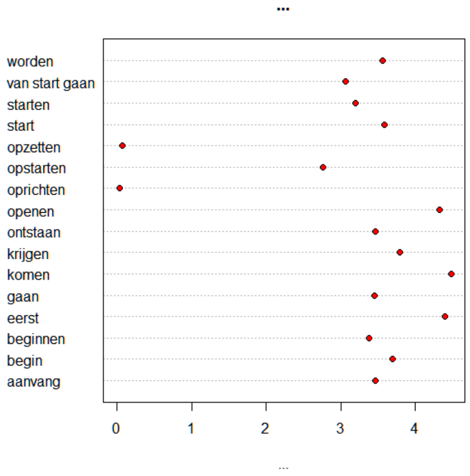


Figure 55 Cluster n°1 for SourceDutch

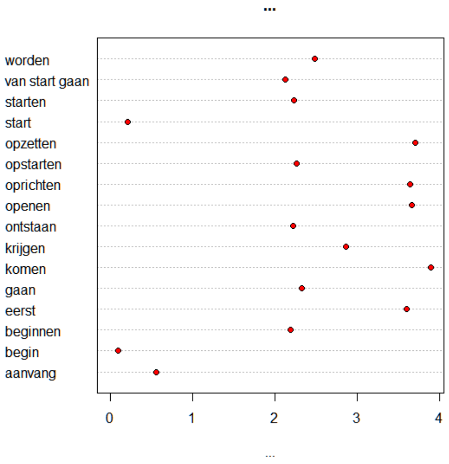


Figure 56 Cluster n°2 for SourceDutch

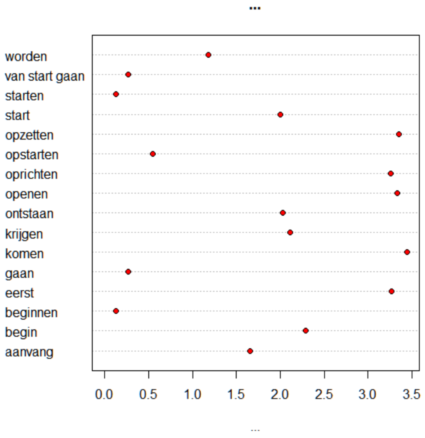


Figure 57 Cluster n°3 for SourceDutch

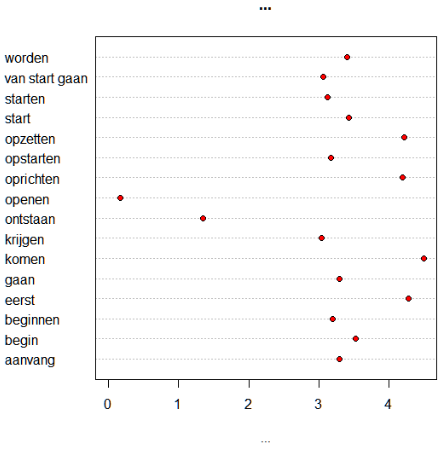


Figure 58 Cluster n°4 for SourceDutch

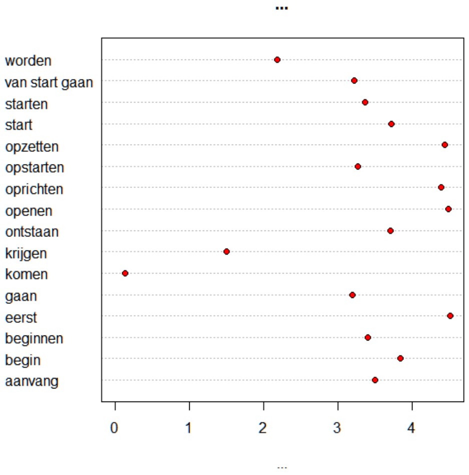


Figure 59 Cluster n°5 for SourceDutch

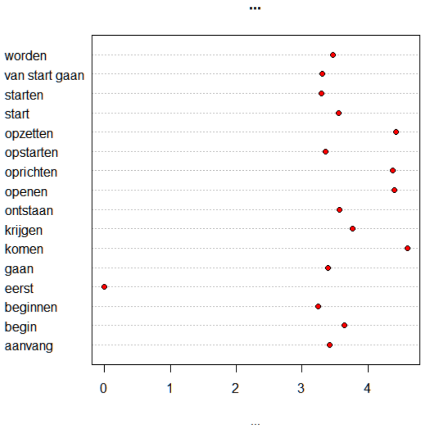


Figure 60 Cluster n°6 for SourceDutch

We observe that the difference in distance of the members of a same cluster to their cluster's *centroid* is often minimal. We therefore use the calculated distances (which are represented by the dots in the dot charts) to evaluate the distances to the *centroids* (see appendix 8).

For cluster n°1, the distance from *opzetten* to the *centroid* is `0.06749455`, whereas the distance from *oprichten* to the *centroid* is `0.02952887`; implying that *oprichten* is closer to the *centroid*, and can hence be considered as the best representation of the abstract prototype of cluster n°1. For cluster n°2, the distance from *start* to the *centroid* is `0.20740218` and the distance from *begin* to the *centroid* is `0.08908994`. This shows that *begin* is closer to the *centroid* and can be indicated as the best representation of the abstract prototype of cluster n°2. For cluster n°3, we see that four lexemes are very close to the zero point. With the bare eye, we can see that *van start gaan*, *gaan* and *opstarten* are slightly further away from the cluster's *centroid*, but the difference in distance between *starten* and *beginnen* is minimal. The distance from *starten* to the *centroid* is `0.1264550`, and the distance from *beginnen* to the *centroid* is `0.1254173`. Hence, *beginnen* is indicated as the cluster's best representation of the abstraction of the prototype. With regard to cluster n°4, Figure 58 clearly shows that *openen* is the closest lexeme to the *centroid*, and can hence be considered as the best representation of the abstract prototype of this cluster. As for cluster n°5, *komen* can clearly be distinguished as the closest lexeme to the *centroid*, and is indicated as its best representation of the abstract prototype. Finally, it is unnecessary to indicate the best representation of the abstract prototype for cluster n°6, which is a singleton cluster with *eerst*.

4.2.3.2 Medoids

A second quantitative possibility to obtain more information about the organization of the lexemes within each cluster is to calculate its *medoid*. The *medoid* assigns *one* object in the cluster from which the average distance to all other objects is the smallest (Divjak 2010, 164). The *medoids* for the clusters are summarized in Table 14, and compared to the lexeme closest to the *centroid* as determined above. As we have explained in section 3.6.3.3, both the lexeme closest to the *centroid* and the *medoid* can be seen as the most prototypical member of the cluster they belong to, but they represent different views on prototypicality.

If we disregard clusters n°1, 4 and 6 – which are clusters with only two or one member – we see that the *medoid* and the lexeme closest to the *centroid* never converge.

Table 14 Comparison of medoids and lexemes closest to the centroids for SourceDutch

	Medoid	**Lexeme closest to *centroids***
Cluster n°2	Start	Begin
Cluster n°3	Starten	Beginnen

Cluster n°5	Krijgen	Komen

All in all, if we assess the results of the calculation of both the *medoids* as well as the distances to the *centroids*, we remark a divergence between the closest lexeme to the *centroid* and the *medoid* of a cluster for all clusters. This divergence increases the uncertainty about which lexeme can be considered as the most central one. In addition, we observe that the difference in distance to the *centroid* is minimal for some clusters, especially for cluster n°3 (*beginnen* vs *starten*), cluster n°2 (*begin* vs *start*). It is noteworthy that for those two clusters with a minimal difference in distance to the *centroid*, it is each time the second closest lexeme that is indicated as the *medoid*. This is potentially very interesting and could indicate a field of tension between several of the more central expressions in each cluster.

These two measures give us important information to help us determine each of the clusters' prototypes and enhance our understanding of the rendered dendrograms. For clusters n°2 and 3, for example, the calculation of the *centroids* and the *medoids* has revealed a 'competition' between *start* and *begin* and *starten* and *beginnen*. The diverging evidence from *medoids* and distance to *centroids* makes it difficult to put forward the outcome of the one or the other measure as the better one to determine the most prototypical expression for each cluster, all the more because we have linked them to different views on prototype. As a consequence (and as we foresaw in section 3.6.3.4) it seems difficult to select the lexeme closest to the *centroid* or the *medoid* as a meta-label to name the specific meaning distinction of the cluster.

4.2.4 Interpretation of the semantic field of *beginnen* / inchoativity for SourceDutch

We will now provide an interpretation of the visualization representing a semantic field of *beginnen* / inchoativity for SourceDutch. 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 meta-labels that we will assign should be understood as a post-hoc, interpretative tool, applied to enhance our understanding of the rendered dendrograms. Note that we do not consider the meta-labels as a 'validation' of the discerned cluster organization – if this had been our intention, we should have determined the labels beforehand. As determined in section 3.6.3.4, information from three types of sources will be used (in addition to the information about the prototype-based organization of the clusters in the field and the lexemes in each cluster): (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; 2013).

We consider cluster n°3 as the most central cluster or REFERENCE CLUSTER, representing the idea of GENERAL ONSET. There are two arguments to justify this. First, on the semasiological level (and as we have seen in section 4.2.2) this cluster's *centroid* is the closest one to the origin of the semantic space and hence, the most central one in the prototype-based organization of the semantic field. Second, on the onomasiological level, if we look at Figures 55 to 60 – which depict the distances of the lexemes to each of the *centroids* of the *other* clusters – we see that the lexemes of cluster n°3 are always situated at a fairly equal distance of the *centroids* of all the other clusters (somewhat in the middle of each plot). This implies that cluster n°3 shows the least deviation with respect to the other clusters (the lexemes of cluster n°3 are all equally similar to the abstract prototype of each of the other clusters). Third, cluster n°3 holds the initial lexeme *beginnen*, which was selected to initiate our SMM++ retrieval task since we consider *beginnen* as the most prototypical expression of inchoativity (based on corpus frequency and etymological age (see section 3.5)). We can conclude that the cluster containing *beginnen* is the one holding the most prototypical expressions of inchoativity.

Let us now take a closer look at the relationships between the lexemes in cluster n°3. 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 & 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 & Gries 2009, 279, our emphasis).

The subdivision observed in our (Dutch) results into verbs formally related to *starten* [to start] such as *van start gaan* [to take off] on the one hand (hence: ACTION verbs), and verbs formally related to *beginnen* [to begin] (hence: STATE AFTER ONSET verbs) on the other hand, thus corroborates the distinction made by Divjak & Gries. The attested distinction between *to start* and *to begin* seems to hold for Dutch *starten* and *beginnen* too. If we look back at the distance from the lexemes to the *centroid* of this cluster, we see that the two lexemes closest to the *centroid* are indeed *beginnen* (`0.1254173`) and *starten* (`0.1264550`); the minimal difference in distance to the *centroid* between these two lexemes further shows that there is some kind of 'competition' going on between the two and that either of the two would be a good candidate to be the best representation of the abstract concept of the prototype. Further note that the distinction between ACTION and STATE AFTER ONSET is not indicated in Cornetto, where all lexemes of cluster n°3 are considered as the same semantic type, i.e. 'action' ("verb that describes an action that is usually controlled by the subject of the verb"), with the only exception that *beginnen* can also be granted the semantic type 'process' ("a dynamic event that is not initiated by an actor capable of acting with volition"). *Gaan* [to go][50], which somewhat oddly seems to be clustered with *beginnen*, is, according to the lexical-semantic database Cornetto (Vossen et al. 2008), defined 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 seems thus 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 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) begin to function], which all bear *beginnen* in their Dutch definition. In sum, we decide to assign the label of **REFERENCE CLUSTER / GENERAL ONSET** to cluster n°3, with REFERENCE CLUSTER referring to the cluster's position in the cluster hierarchy and GENERAL ONSET representing the overall semantic content of this cluster. 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 the prototypical verbs *beginnen* and *starten* – as well as *aanvang*. On the semasiological level, we see that the *centroid* of this cluster is the second closest one to the zero-point, implying its relative centrality in the semantic space. The *centroid* of cluster n°2 is also fairly close to the *centroid* of the REFERENCE CLUSTER, which seems to confirm the close relationship between the two clusters. The third lexeme in this cluster, *aanvang* is again a noun, but differs from *begin* and *start* in that it belongs to a more formal register (Van

[50] Recall that observations of *gaan* in the construction *van start gaan* are not included here.

Dale 2015). 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*[51] (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 meta-label for cluster n°2. With GENERAL ONSET we indicate that this cluster situates itself close to the REFERENCE CLUSTER of GENERAL ONSET; the addition of (NOUN) refers to the word-class dependence of this cluster.

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 (10 and 11) from the DPC show that *oprichten* can, just as *opzetten*, be used in business-like contexts:

(10) In 2000 **zetten** de twee bedrijven een joint venture **op** in Turkije. Vandaag doen zij dat opnieuw in Roemenië. [SOURCE: In 2000 the two companies **set up** a joint venture together in Turkey and today they are launching another in Romania]. (dpc-arc-002048-en).

(11) 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).

On the onomasiological level, the difference in distance of the two lexemes to their cluster's *centroid* was very small. Although *oprichten* (0.02952887) was situated slightly closer to the *centroid*, *opzetten* (0.06749455) was indicated as the *medoid*. This information further substantiates the idea that *oprichten* and *opzetten* are indeed near-synonyms. 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.

[51] but not *word choice*

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) as the following examples shows (note that, in this example (12), the inchoative aspect is explicitated by its translation):

(12) '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*, viz., "beginnen te spreken" [start to speak], of *krijgen*, viz., "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 13) shows that *openen* can, just as *ontstaan* refer to abstract processes, such as the coming into being of a right:

(13) 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 particularity of *openen* in this field is that its inchoative meaning is in fact a metaphorical meaning extension of its clear literal meaning ("to open a door, a window"). "To open a new business unit" indicates that a new business unit is set up/comes into being, as illustrated in example 14 below:

(14) In het kader van de concentrische groei,[...], *opende* men een Nederlandse distributieafdeling in Tilburg. [TARGET Within the framework of concentric growth, [...], a Dutch distribution department was set up in Tilburg]. (dpc-lan-001674-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.

In sum, we labeled the different meaning distinctions (clusters) within the semantic field of *beginnen* / inchoativity as follows (see Figure 61): cluster n°3 (*opstarten* [to start up], *starten* [to start], *van start gaan* [to take off], *beginnen* [to begin] and *gaan* [to go]) is labeled as **REFERENCE CLUSTER / GENERAL ONSET**. Within cluster n°3, we have furthermore discerned an additional meaning distinction between *beginnen* [to begin], *gaan* [to go] labeled as **STATE AFTER ONSET** and *starten* [to start], *van start gaan* [to take off] labeled as **ACTION**. Cluster n°2 (*aanvang* [commencement], *begin* [beginning] and *start*[start]) is labeled as **GENERAL ONSET (NOUN)**, cluster n°1 (*oprichten* [to establish] and *opzetten* [to set up]) received the label **SPECIFIC ACTION**, cluster n°5 (*komen* [to come], *krijgen* [to get] and *worden* [to become]) is labeled as **NON-LEXICALIZED INCHOATIVITY**. Cluster n°4 (*ontstaan* [to come into being] and *openen* [to open]) is labeled as **ONSET OF ABSTRACT PROCESSES**. Obviously these meta-labels are far from ideal descriptions of the clusters and are naturally open for discussion. As announced in the introduction of this chapter, the meta-labels merely serve to enhance our understanding of the clusters and to facilitate the further description of what happens to the meaning distinctions revealed by the clusters in the different semantic fields.

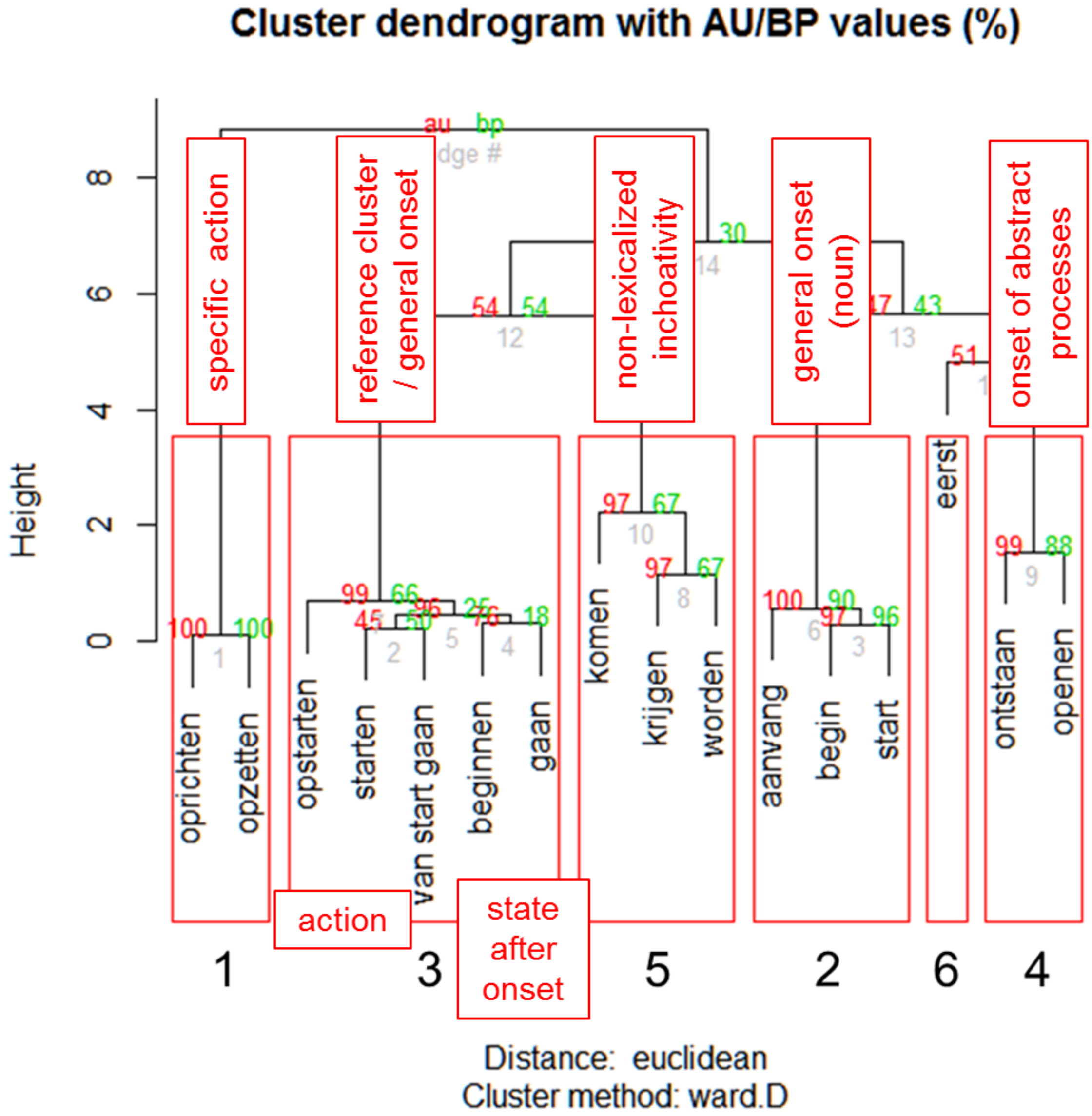


Figure 61 Dendrogram representing a semantic field of *beginnen* / inchoativity for SourceDutch with meta-labels

4.3 TransDutch$_{ENG}$

For the description and interpretation of TransDutch$_{ENG}$, we repeat the same steps as for SourceDutch, presented in the previous section.

4.3.1 Results of the Hierarchical Agglomerative Cluster analysis

The distribution of the variation over the latent dimensions of the CA is shown in Figure 62 and Figure 63. We choose to reduce the number of dimensions of the CA to 4[52].

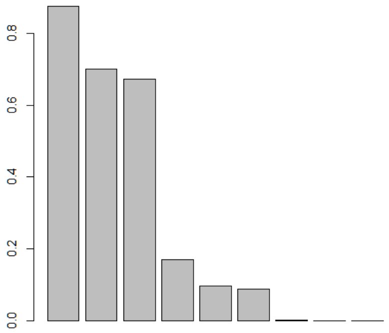


Figure 62 Scree plot for TransDutch$_{ENG}$

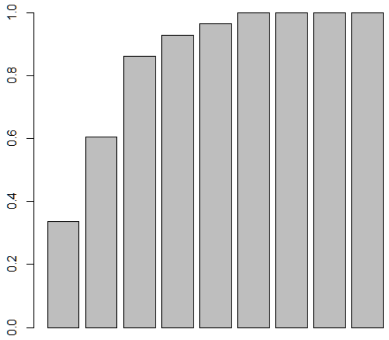


Figure 63 Cumulative scree plot for TransDutch$_{ENG}$

A HAC is now carried out on the output of the CA. The cut-off point is set at a height of 2, which offers a cluster solution with 6 clusters[53]. Cluster n°1 contains *oprichten* [to establish] and *opzetten* [to set up]; cluster n°2 includes *aanvang* [commencement] and *start* [start]; cluster n°3 comprises *eerst* [firstly], *van start gaan* [to take off], *beginnen* [to begin], *krijgen* [to get], *starten* [to start], *gaan* [to go], *worden* [to become]; cluster n°4

[52] Although 3 dimensions would seem to suffice here to represent more than 80% of the variation, we opt for 4 dimensions (which is the minimum number of dimensions required to carry out `pvclust()` in the next step of this analysis).

[53] Note that – had we applied `pvrect()`, which cuts off each cluster at the highest possible node with a significant p-value – the same cluster solution would have been obtained.

holds *komen* [to come] and *opstarten* [to start up], cluster n°5 consists of *ontstaan* [to come into being] and *openen* [to open] and cluster n°6 contains *begin* [beginning]. We consider the result presented in Figure 64 as a possible visualization of a semantic field representing *beginnen* / inchoativity in TransDutch$_{\text{ENG}}$[54].

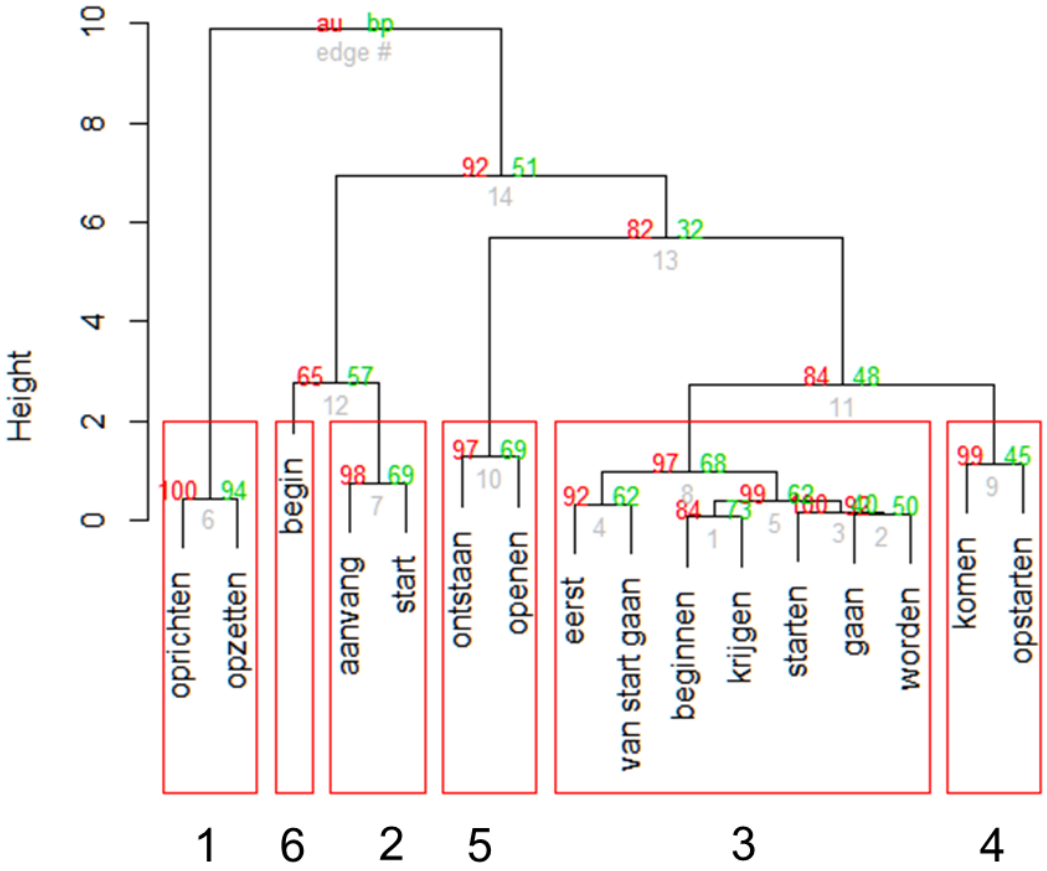


Figure 64 Dendrogram representing a semantic field of *beginnen* / inchoativity for TransDutch$_{\text{ENG}}$

The chosen cluster solution is validated on the basis of the *average silhouette width*. For a solution with 6 clusters for TransDutch$_{\text{ENG}}$ we obtain an *average silhouette width* of `0.57`, which we consider to indicate a good classification.

[54] For the reasoning behind the assignment of the numerals to the clusters, please refer to section 4.3.2.

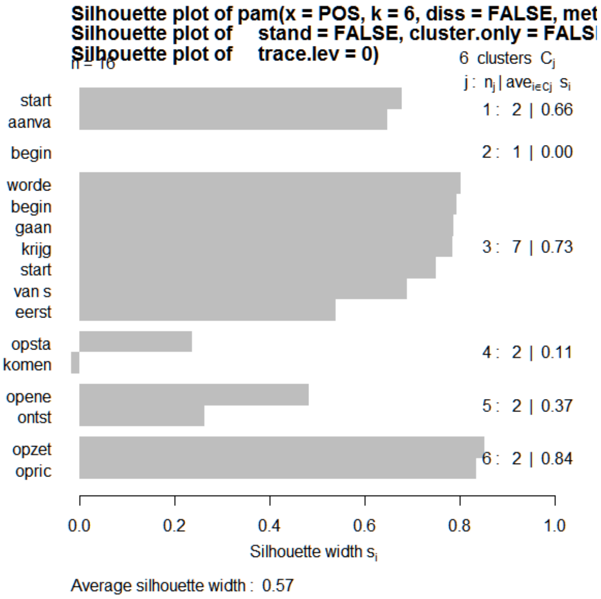


Figure 65 Average silhouette width for cluster solution with 6 clusters for TransDutch$_{ENG}$

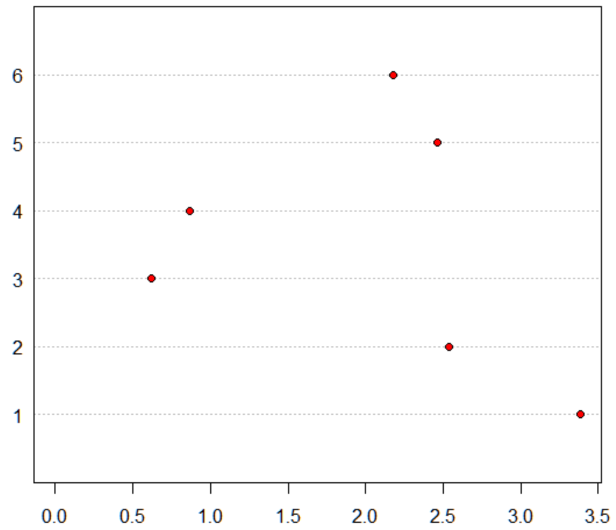
A second validation is obtained via the calculation of a K-means clustering. When a cluster solution with 6 clusters is requested, K-means proposes the following solution (the numeral beneath each lexeme assigns it to a specific cluster):

```
Clustering vector:
       aanvang          begin       beginnen          eerst           gaan
             1              2              3              3              3
         komen        krijgen       ontstaan         openen      oprichten
             4              3              5              5              6
     opstarten       opzetten          start        starten van start gaan
             4              6              1              3              3
        worden
             3
```

The cluster solution proposed by the K-means clustering with 6 clusters is identical to the output of the HAC. On the basis of both validation techniques, we can conclude that the chosen cluster solution for TransDutch$_{ENG}$ is a good classification.

4.3.2 Prototype-based organization of the clusters in the dendrogram (semasiological level)

We calculate the *centroid* of each cluster and assess its distance to the zero-point of the semantic space by mapping the *centroids* onto a dot chart (Figure 66). The content of each cluster number in the dot chart is summarized in the table accompanying Figure 66[55]:



Cluster 6	begin
Cluster 5	ontstaan, openen
Cluster 4	komen, opstarten
Cluster 3	beginnen, eerst, gaan, krijgen, starten, van start gaan, worden
Cluster 2	aanvang, start
Cluster 1	oprichten, opzetten

Figure 66 Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of TransDutch$_{ENG}$

The dot chart shows us that cluster n°3, containing *beginnen*, *eerst*, *gaan*, *krijgen*, *starten*, *van start gaan* and *worden* is the central cluster in the analysis, closely followed by cluster n°4 with *komen* and *opstarten*. Clusters n°6 with *begin*, n°5 with *ontstaan* and *openen* and n°2 with *aanvang* and *start* are situated closely together, but further away from the plot's origin. Cluster n°1 comprising *oprichten* and *opzetten* is the most peripheral cluster.

4.3.3 Prototype-based organization of the lexemes within each cluster (onomasiological level)

The prototype-based organization of the lexemes within each cluster is examined by measuring the distance of the lexemes within each cluster to the *centroid* of the cluster they belong to. We also calculate the *medoid* of each cluster. Both measures will be used to determine which lexical item can be considered as the most prototypical expression of the cluster it belongs to.

[55] Parallel to SourceDutch, the numerals on the y-axis of the dot chart in Figure 66 are assigned by a previously established list (based on the output of the cluster analysis), necessary to calculate the cluster *centroid* (the order of the assigned numerals is arbitrary).

4.3.3.1 Centroids

The dot charts in Figures 67 to 72 represent the distance of all the lexemes to the *centroid* (the abstraction of the prototype) of a particular cluster.

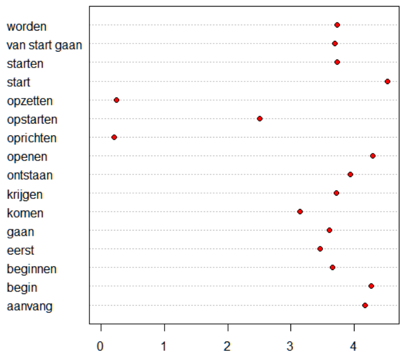


Figure 67 Cluster n°1 for TransDutch$_{\text{ENG}}$

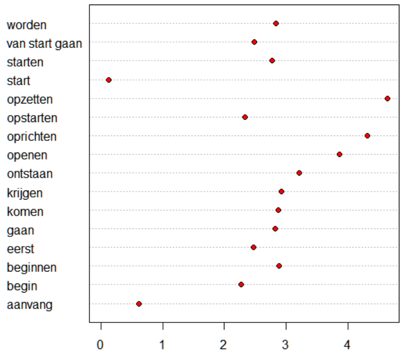


Figure 68 Cluster n°2 for TransDutch$_{\text{ENG}}$

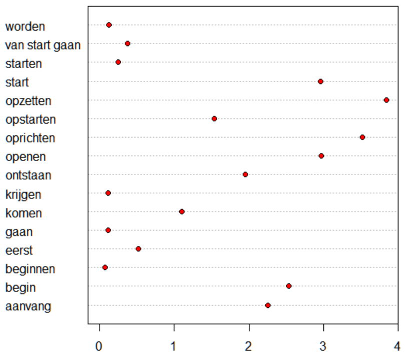


Figure 69 Cluster n°3 for TransDutch$_{\text{ENG}}$

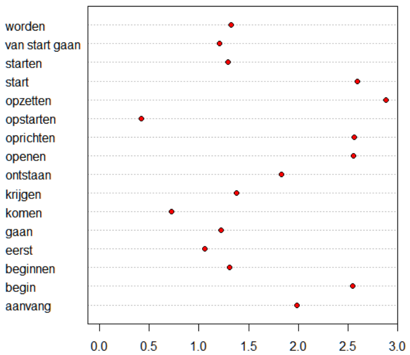


Figure 70 Cluster n°4 for TransDutch$_{\text{ENG}}$

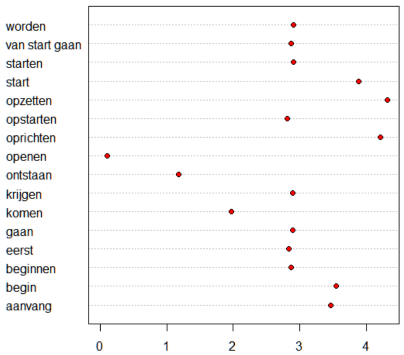


Figure 71 Cluster n°5 for TransDutch$_{\text{ENG}}$

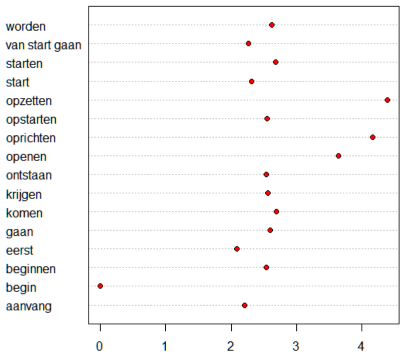


Figure 72 Cluster n°6 for TransDutch$_{\text{ENG}}$

Just as for SourceDutch, the differences in distance of the lexemes to their cluster's *centroid* is often very small, so we will again use the calculated distances whenever the dot charts do not clearly indicate which lexeme is the closest one to the *centroid* (see appendix 9).

When we look at the calculated distances for cluster n°1, we see that *oprichten* is slightly closer to the cluster's *centroid* (`0.2004520`) than *opzetten* (`0.2476172`). As for cluster n°2, we can see that *start* is the lexeme closest to the *centroid* of the cluster. In cluster n°3, *beginnen* (`0.06521312`) is closer to the *centroid* than *gaan* (`0.11345029`), *krijgen* (`0.11370695`) and *worden* (`0.12738579`). For cluster n°4, *opstarten* is undoubtedly the closest lexeme to the *centroids* of the cluster. Also note that the second lexeme in cluster n°4, *komen* is situated as close to *opstarten* (of cluster n°4) as it is to *eerst* (of cluster n°3), and also quite close to a number of other lexemes pertaining to cluster n°3. This implies that the clustering of *komen* with *opstarten* is not so clear cut. Looking back at cluster n°3, we indeed see that *komen* is the lexeme that is situated closest to the lexemes of cluster n°3. For cluster n°5, it is *openen* which situates itself closest to the cluster *centroids*. For cluster n°6, there is no need to determine the best representation of the abstraction of the prototype since we are dealing here with a singleton cluster with *begin*.

4.3.3.2 Medoids

Table 15 below shows the calculated *medoid* for cluster n°3 and compares it with the lexemes closest to the *centroid* of the cluster (all other clusters contain either two lexemes or only one, so the *medoid* could not be calculated).

Table 15 Comparison of medoids and lexemes closest to the centroids for TransDutch$_{ENG}$

	Medoid	**Lexeme closest to *centroids***
Cluster n°3	worden	beginnen

We again see that the *medoid* and the closest lexeme to the *centroid* of cluster n°3 do not coincide. Second, the difference in distance to the *centroid* between the first and the second lexeme points to a lesser extent than in SourceDutch towards the presumed 'competition' between several more central expressions within the cluster (recall that for SourceDutch, the second closest lexeme to the *centroid* was always designated as the *medoid*): for cluster n°3, *beginnen* is now closely followed by *gaan*, *krijgen* and *worden*. *Starten* – for which we would have expected a more central position in the cluster– is situated slightly further away.

4.3.4 Interpretation of the semantic field of *beginnen* / inchoativity for TransDutch$_{ENG}$

We now provide an interpretation – which includes the assignment of a meta-label for each meaning distinction – of a semantic field of *beginnen* / inchoativity for TransDutch$_{ENG}$. The specific meaning distinctions determined for SourceDutch will be used as a point of reference to interpret the field of TransDutch$_{ENG.}$ We will consequently 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 most central cluster or REFERENCE CLUSTER, representing the idea of GENERAL ONSET. Parallel to SourceDutch, this is substantiated on both the semasiological and the onomasiological level. On the semasiological level, we see that the *centroid* of cluster n°3 is the closest one to the origin of the semantic space (considered as the prototypical center). On the onomasiological level, we observe that the distances of the lexemes of cluster n°3 to each of the *centroids* of the *other* clusters (depicted in Figures 67 to 72) are always fairly equal (with the exception of cluster n°4). This implies that cluster n°3 shows the least deviation with respect to the other clusters (equally similar to the abstract prototype of each of the other clusters). Within the REFERENCE CLUSTER, we furthermore find the initial lexeme *beginnen* (considered as the most prototypical expression of inchoativity), strengthening our assumption that this cluster is holding the most prototypical expressions of inchoativity. We notice that the REFERENCE CLUSTER has become larger compared to SourceDutch: *eerst* – which held a peripheral position in SourceDutch (outliers are often depicted as singleton clusters in a HAC) – is now part of the REFERENCE CLUSTER, 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 sub-nodes of the REFERENCE CLUSTER, 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 REFERENCE CLUSTER 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*). At first sight, it seems that within the REFERENCE CLUSTER of TransDutch$_{ENG}$, the emphasis is on the wider relatedness between the verbs rather than on the division between ACTION and STATE AFTER ONSET. However, if we take a look at the onomasiological level by assessing the distance from each of the lexemes to the *centroid* of the cluster, we see that *beginnen*

(`0.06521312`) is the closest lexeme to the *centroid*, followed by *gaan* (`0.11345029`), which is considered as a STATE AFTER ONSET verb, followed by two verbs labeled as NON-LEXICALIZED INCHOATIVITY, i.e. *krijgen* (`0.11370695`) and *worden* (`0.12738579`); followed by the ACTION verbs *starten* (`0.25003812`) and *van start gaan* (`0.37259612`). Seen from this perspective, the 'confusion' of ACTION and STATE AFTER ONSET verbs within the REFERENCE CLUSTER is much less present than the dendrogram would seem to suggest. We could rather state that in TransDutch$_{ENG}$, the competition between ACTION and STATE AFTER ONSET verbs has been breached by a more prominent use of verbs which do not lexicalize inchoativity.

Cluster n°4 is a somewhat odd, new cluster. From the dot chart in section 4.3.2, we know that this cluster is the closest one to the REFERENCE CLUSTER, confirming its close relatedness with the latter. Since the REFERENCE CLUSTER contains the ACTION verbs as well as verbs of NON-LEXICALIZED INCHOATIVITY, one would have expected *opstarten* and *komen* in the REFERENCE CLUSTER too. There are indeed a number of indications that cluster n°4 is very closely related to the REFERENCE CLUSTER: (i) the lexemes of cluster n°4 seem to behave in a similar way to those of cluster n°3: the lexemes of both clusters keep a similar distance from the *centroids* of the other clusters, implying that they show very little deviation with respect to the other clusters (and the same amount of deviation for both clusters n°3 and n°4); (ii) if we furthermore inspect the distance of the lexemes *komen* and *opstarten* to the lexemes of the REFERENCE CLUSTER (Figure 70), we see that *komen* (`0.7203757`) is as close to *eerst* (`1.0569324`) as it is to *opstarten* (`0.4202192`). Hence, it is mainly *opstarten* that determines the separate clustering here (*komen* holds a middle position between clusters n°3 and n°4). Recall that in SourceDutch, *opstarten* already formed a significant sub-node within the REFERENCE CLUSTER. 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. On the semasiological level, if we assess the distance of the *centroid* of cluster n°2 and cluster n°6 to the zero-point of the semantic space, we see that cluster n°6 (*begin*) is much closer to the zero-point than cluster n°2, implying that in TransDutch$_{ENG}$, *begin* is a more central expression of inchoativity than *aanvang* and *start* are. In TransDutch$_{ENG}$, the distance between *aanvang* and *start* is also larger (Figure 68) compared to SourceDutch (Figure 56).

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. On the onomasiological level, we do see that the difference in distance to the centroid of the lexemes of cluster n°1 (*oprichten* and *opzetten*) has become larger in TransDutch$_{ENG}$, compared to the corresponding cluster in SourceDutch. For cluster no°5, (*ontstaan* and *openen*) the difference in distance to the *centroid* has become smaller in TransDutch$_{\text{ENG}}$ compared to SourceDutch. Figure 73 below now shows the semantic field of *beginnen* / inchoativity for TransDutch$_{\text{ENG}}$ with the meta-labels.

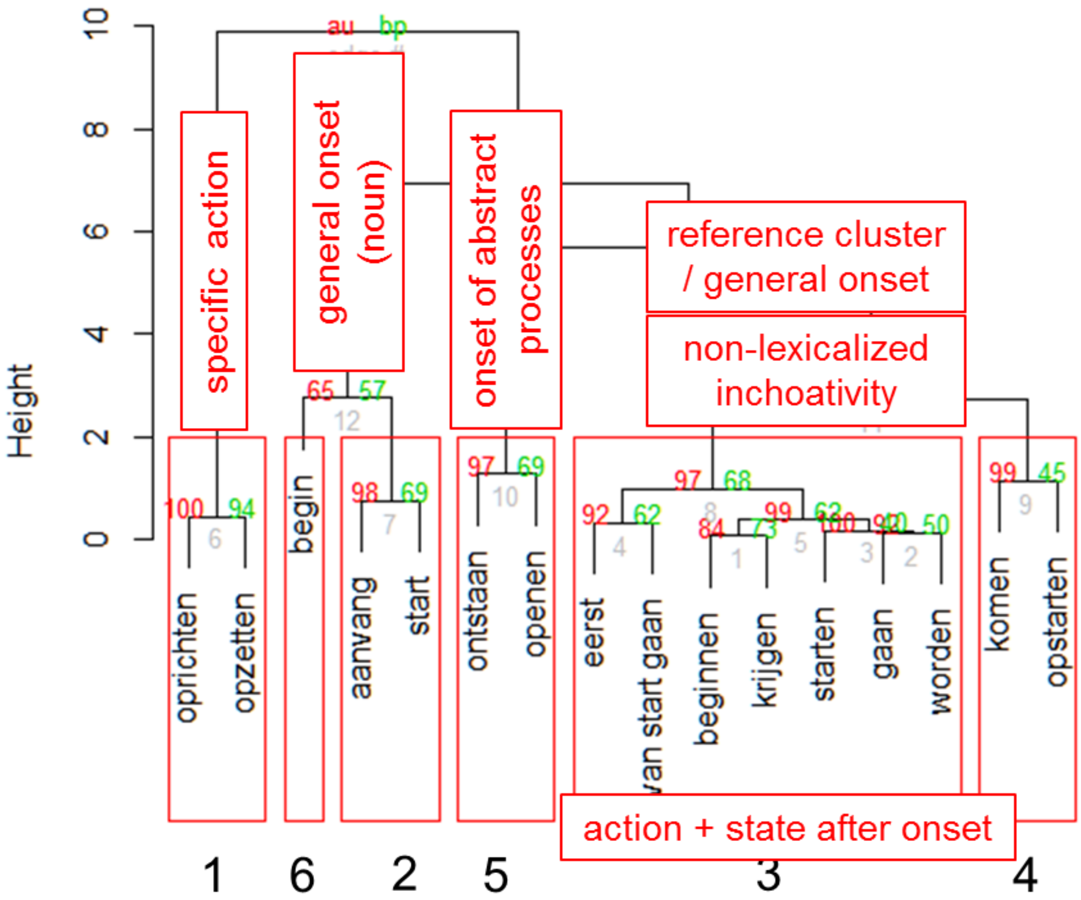


Figure 73 Dendrogram representing a semantic field of *beginnen* / inchoativity for TransDutch$_{ENG}$ with meta-labels

4.4 TransDutch$_{FR}$

In this section, we present our interpretation of the visualization of TransDutch$_{FR}$ following the same steps as for SourceDutch and TransDutch$_{ENG}$.

4.4.1 Results of the Hierarchical Agglomerative Cluster analysis

Figures 74 and 75 show the distribution of the variation over the latent dimensions of the CA. On the basis of these scree plots, it is decided to reduce the number of dimensions of the CA to 4.

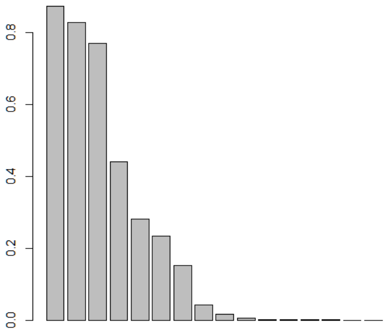


Figure 74 Scree plot for TransDutch$_{FR}$

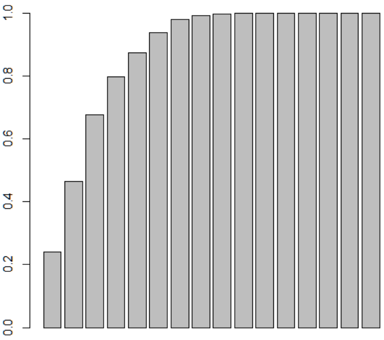


Figure 75 Cumulative scree plot for TransDutch$_{FR}$

We now carry out a HAC and choose a cut-off point at a height of 5, rendering a cluster solution with 4 clusters. Cluster n°1 contains *start* [start], *aanvang* [commencement] and *begin* [beginning]; cluster n°2 includes *ontstaan* [to come into being] and *openen* [to open]; cluster n°3 comprises *opzetten* [to set up], *oprichten* [to establish], *opstarten* [to start up], *starten* [to start] and *van start gaan* [to take off]; cluster n°4 holds *eerst* [firstly], *gaan* [to go], *beginnen* [to begin], *worden* [to become], *komen* [to come] and *krijgen* [to get].

We consider the result presented in Figure 76 as a possible visualization of a semantic field representing *beginnen* / inchoativity in TransDutch$_{FR}$[56].

[56] For the reasoning behind the assignment of the numerals to the clusters, please refer to section 4.4.2.

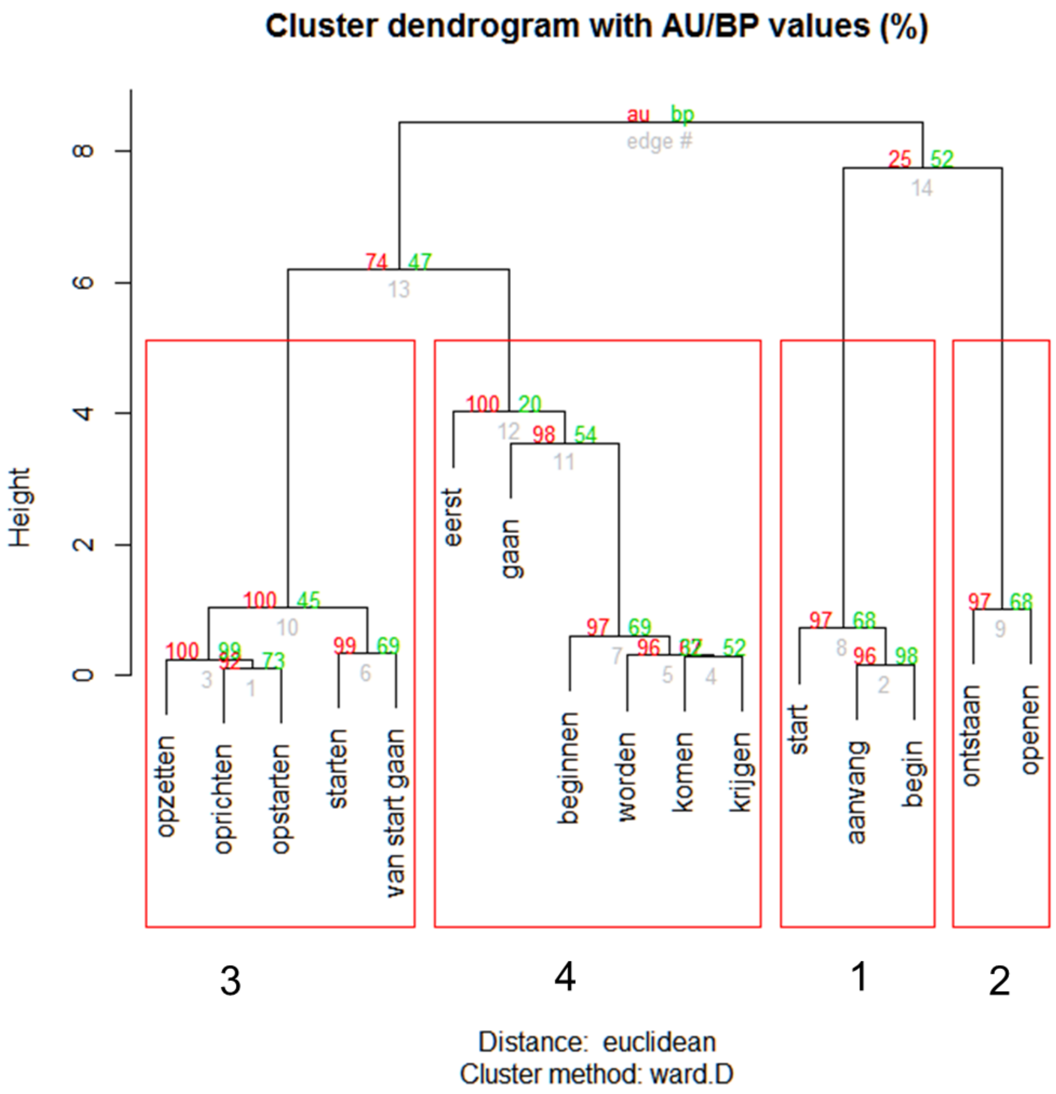


Figure 76 Dendrogram representing a semantic field of *beginnen* for TransDutch$_{FR}$

Our cluster solution is validated by the *average silhouette width* for a solution with 4 clusters (*average silhouette width* = `0.53`) (Figure 77) and by the calculation of a K-means clusters with 4 clusters, which proposes an identical cluster solution to the output of the HAC as can be seen below (the numeral beneath each lexeme assigns it to a specific cluster). On the basis of both validation techniques, we conclude that the chosen cluster solution for TransDutch$_{FR}$ can be considered a good classification.

```
Clustering vector:
        aanvang            begin        beginnen            eerst             gaan
              1                1               4                4                4
         komen          krijgen        ontstaan           openen        oprichten
              4                4               2                2                3
      opstarten         opzetten           start  starten van start gaan
              3                3               1                3                3
         worden
              4
```

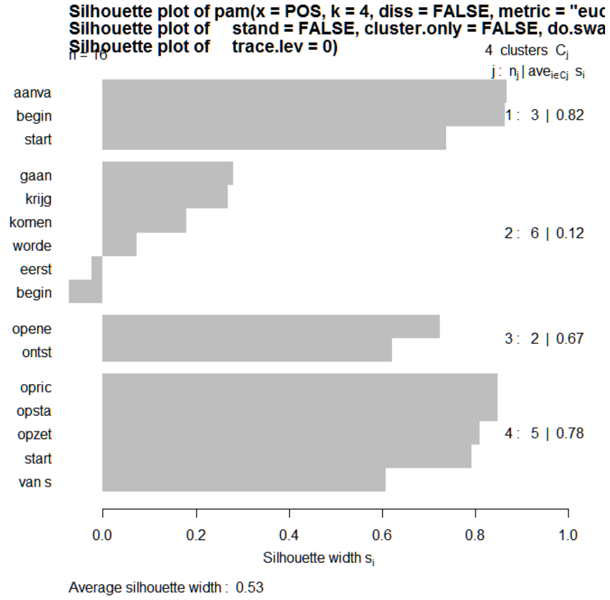
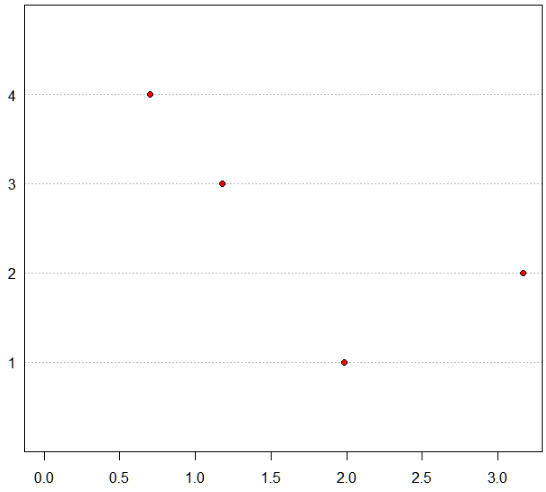


Figure 77 Average silhouette width for cluster solution with 4 clusters for TransDutch$_{FR}$

4.4.2 Prototype-based organization of the clusters in the dendrogram (semasiological level)

The distance from each cluster's *centroid* to the zero-point of the semantic space is calculated and mapped on a dot chart (Figure 78). The content of each cluster number in the dot chart is summarized in the table accompanying Figure 78:



Cluster 4	eerst, gaan, beginnen, komen, worden, krijgen
Cluster 3	opzetten, oprichten, opstarten, starten, van start gaan
Cluster 2	ontstaan, openen
Cluster 1	start, aanvang, begin

Figure 78 Dot chart presenting the distance of the cluster centroids to the zero-point of the semantic space of TransDutch$_{FR}$

Cluster n°4, containing *eerst*, *gaan*, *beginnen*, *komen*, *worden*, *krijgen* is the central cluster in the analysis since it is situated closest to the zero-point of the semantic space. Cluster n°3 with *opzetten*, *oprichten*, *opstarten*, *starten* and *van start gaan* comes in second place and is followed by cluster n°1 (*start*, *aanvang*, *begin*). The cluster that is furthest away from the zero-point of the semantic space is cluster n°2 comprising *ontstaan* and *openen*.

4.4.3 Prototype-based organization of the lexemes within each cluster (onomasiological level)

The prototype-based organization of the lexemes within each cluster is examined on the basis of the following measures. We calculate the distance of the lexemes within each cluster to the *centroid* of the cluster they belong to and we calculate the *medoid* of each cluster. Both measures can also be used to determine which lexical item can be considered as the most prototypical expression of the cluster it belongs to.

4.4.3.1 Centroids

The dot charts in Figures 79 to 82 represent the distances of all the lexemes in the analysis to the *centroid* of one particular cluster. The lexeme closest to the cluster's *centroid* can be considered as the best representation of the abstract idea of the prototype of that cluster. We again use the calculated distances (which are represented by the dots in the dot charts) to evaluate the distances to the *centroids* (see appendix 10).

For cluster n°1, *begin* is the closest lexeme to the *centroid*, situated at `0.05884857` of the *centroid*, followed by *aanvang* at `0.12955053` and *start* at `0.54160901` of the *centroid*. For cluster n°2, it is clear that *openen* is the lexeme closest to the *centroid* of its cluster. As for cluster n°3, it is difficult to determine with the bare eye whether *starten* (`0.1160007`) or *oprichten* (`0.2037736`) is the lexeme closest to the *centroid*, but based on the calculated distances, we can conclude that *starten* is the closest one to the *centroid* of the cluster. Finally, for cluster n°4, we see that *beginnen* is the lexeme closest to the cluster's *centroid* (`0.6576414`), followed by *krijgen* (`0.9121243`). It is worthy to note here that the closest lexeme to the REFERENCE CLUSTER, *beginnen*, is situated at a relatively large distance of its cluster's *centroid* (`0.6576414).` When we compare the distance of *beginnen* to the *centroid* of the REFERENCE CLUSTER it belongs to for SourceDutch (`0.1254173`) and TransDutch$_{ENG}$ (`0.06521312`), we see that *beginnen* is closest to its *centroid* (the abstract prototype) for TransDutch$_{ENG}$ and furthest for TransDutch$_{FR}$.

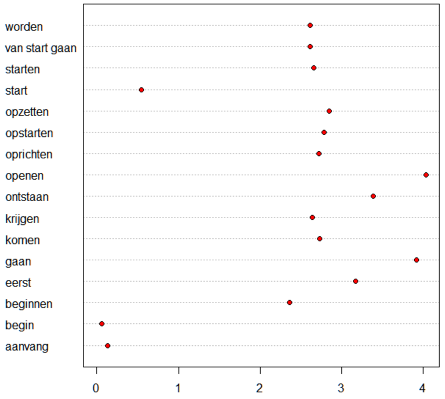


Figure 79 Cluster n°1 for TransDutch$_{FR}$

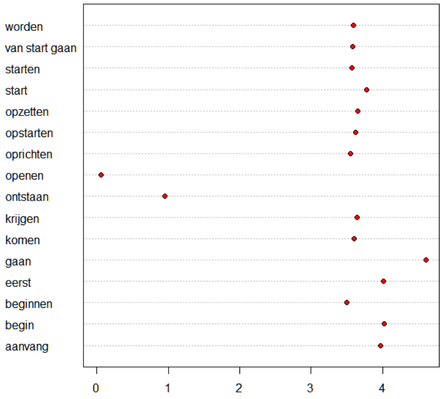


Figure 80 Cluster n°2 for TransDutch$_{FR}$

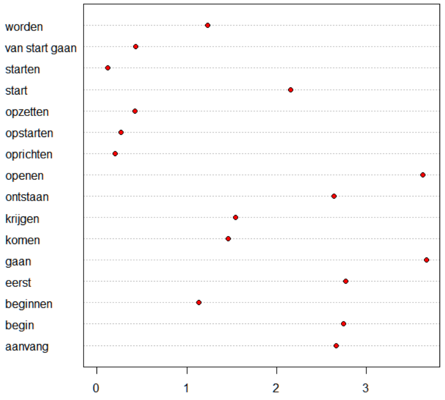


Figure 81 Cluster n°3 for TransDutch$_{FR}$

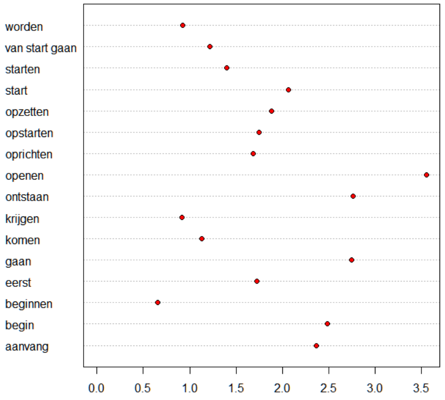


Figure 82 Cluster n°4 for TransDutch$_{F}$

4.4.3.2 Medoids

In Table 16 below, we compare the lexemes closest to the *centroid* of clusters n°1, 3 and 4 to their respective *medoid*.

Table 16 Comparison of medoids and lexemes closest to the centroids for TransDutch$_{FR}$

	Medoid	**Lexeme closest to *centroids***
Cluster n°1	aanvang	begin
Cluster n°3	oprichten	starten
Cluster n°4	krijgen	beginnen

For TransDutch$_{FR}$, the *medoid* and the lexeme closest to the *centroid* never coincide. What is striking is that the *medoid* is each time the second closest lexeme to the *centroid* of the cluster, an observation that was also made for a number of clusters of SourceDutch. Moreover, for clusters n°3 and n°4, we see that their *medoids* indicate one meaning distinction: *oprichten* in cluster n°3 refers to SPECIFIC ACTION and *krijgen* in cluster n°4 refers to NON-LEXICALIZED INCHOATIVITY. For the same clusters, the lexemes closest to the *centroids* indicate a different meaning distinction within the same cluster: ACTION for cluster n°3 (*starten*) and STATE AFTER ONSET for cluster n°4 (*beginnen*).

4.4.4 Interpretation of the semantic field of *beginnen* / inchoativity for TransDutch$_{FR}$

We now provide an interpretation of a semantic field of *beginnen* / inchoativity for TransDutch$_{FR}$. 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 most central cluster in the dendrogram, representing the idea of GENERAL ONSET. As we saw in section 4.4.2, its *centroid* is the closest one to the zero-point of the semantic space, considered as the prototypical center of the semantic space (semasiological level). Just as for SourceDutch and TransDutch$_{ENG}$, *beginnen* is part of the REFERENCE CLUSTER, leading to the assumption that this cluster contains the most prototypical expressions of inchoativity. Parallel to TransDutch$_{ENG}$, the number of lexemes in the REFERENCE CLUSTER has increased compared to SourceDutch (5 lexemes in the REFERENCE CLUSTER of SourceDutch, 7 for TransDutch$_{ENG}$ and 6 for TransDutch$_{FR}$) (onomasiological level). 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 REFERENCE CLUSTER. 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 REFERENCE CLUSTER, we discern two significant terminal nodes (*eerst* and *gaan*), and one significant sub-node with four leaves with *beginnen* as a significant terminal node within the sub-node and a second, underlying sub-node (also significant) with the three verbs labeled as NON-LEXICALIZED INCHOATIVITY. Within this REFERENCE CLUSTER, the meaning distinctions STATE AFTER ONSET and NON-LEXICALIZED INCHOATIVITY are both present. An important difference with SourceDutch and TransDutch$_{ENG}$ is that the REFERENCE CLUSTER of TransDutch$_{FR}$ 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 REFERENCE CLUSTER, and that for TransDutch$_{ENG}$, this distinction was still present in the REFERENCE CLUSTER although less clear (see section 4.2.4).

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 TransDutch$_{FR}$, the distinction between ACTION and STATE AFTER ONSET verbs is marked more clearly, compared to both SourceDutch and TransDutch$_{ENG}$: 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 REFERENCE CLUSTER of SourceDutch and a separate cluster in TransDutch$_{ENG}$) 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, i.e. 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 (15 and 16) from the DPC:

(15) 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)

(16) 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)

On the semasiological level, the *centroid* for cluster n°3 is the second closest one to the zero-point of the semantic space. Its *centroid* is also situated fairly close to the *centroid* of cluster n°4, the REFERENCE CLUSTER, which seems to confirm the close relationship between the two clusters and the proximity of cluster n°3 to the REFERENCE CLUSTER. The proximity between cluster n°3 and cluster n°4 is further confirmed on the onomasiological level. When we look at the distance of the lexemes to the *centroid* of either cluster (Figures 81 and 82), we remark that the image is quite different from what we usually see for the other clusters. In general, the lexemes pertaining to the cluster of which the *centroid* is taken as the zero-point are clearly closer to the *centroid* of their own cluster compared to the other lexemes not pertaining to the cluster. For the lexemes pertaining to clusters n°3 and n°4, the dot charts do not (as) clearly differentiate the lexemes pertaining to their own cluster from those pertaining to the other cluster: a number of lexemes are indeed at a fairly equal distance of both the *centroids* of cluster n°3 and cluster n°4 (see e.g. *komen* is situated at 1.4586546 from the *centroid* of cluster n°3 and at 1.1315485 from the *centroid* of cluster n°4). The close relatedness between clusters n°3 and n°4 is no total surprise since these clusters contain the ACTION verbs in cluster n°3 and the STATE AFTER ONSET verbs in cluster n°4 (which in SourceDutch and TransDutch$_{ENG}$ were separate sub-nodes of their REFERENCE CLUSTERS). Conclusively, the lexemes that were covered under the meta-label REFERENCE CLUSTER/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, i.e. 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. The *centroid* of cluster n°1 is closely following the *centroid* of clusters n°3 and n°4 (Figure 78), confirming the relatedness of this cluster of nouns to the two more central clusters (semasiological level). Note that the only three nouns in the set of lexemes are again clustered together, confirming again the word-class dependent clustering. In addition, when we assess the distance from the lexemes to their cluster's *centroid*, we see that *begin* and *aanvang* are the closest ones to the *centroid*, *start* is situated considerably further away. Although the overall clustering of the three lexemes into one meaning distinction is similar to SourceDutch, the distance from the lexemes to their cluster's *centroids* is different (we observe small differences on the onomasiological level): for SourceDutch, *start* and *begin* are competing to be the closest lexeme to the *centroid*, with *aanvang* situated somewhat further away, whereas in TransDutch$_{FR}$, *aanvang* is much closer to *begin* (the closest lexeme to the *centroid*) and *start* is situated further away. The situation is also very 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$_{ENG}$. The distance from the two lexemes to the *centroids* of their cluster remains also fairly equal throughout the three visualizations. Figure 83 shows the semantic field of *beginnen* for TransDutch$_{FR}$ with integration of the meta-labels.

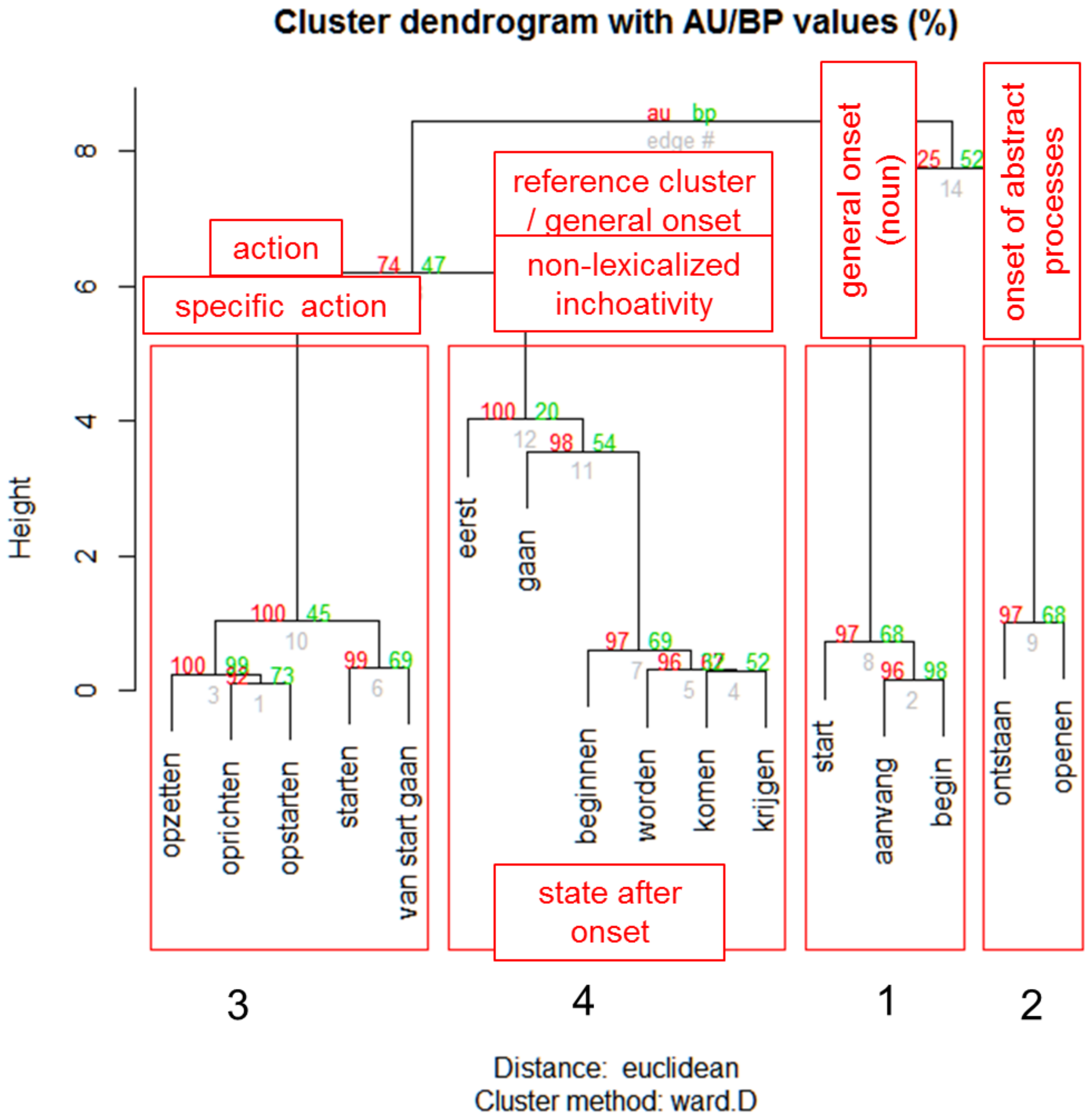


Figure 83 Dendrogram representing a semantic field of *beginnen* for TransDutch$_{FR}$ with meta-labels.

In conclusion, the following similarities have been observed for the three visualizations: For all three visualizations, the cluster closest to the zero-point of the semantic space (considered as the prototypical center) was indicated as the REFERENCE CLUSTER. In addition, the initial lexeme *beginnen* is part of the REFERENCE CLUSTER in all three visualizations.. Since we consider *beginnen* as the most prototypical expression of inchoativity, our belief is that the REFERENCE CLUSTER/GENERAL ONSET contains the most prototypical expressions of inchoativity.

For all three visualizations, the distance of the lexemes in the REFERENCE CLUSTER to the abstract prototypes of the other clusters is fairly equal. This implies that the REFERENCE CLUSTER is indeed the most central one in the semantic space and shows the least deviation with respect to the other clusters (the lexemes in the REFERENCE

CLUSTER are all fairly equally similar to the abstract prototypes of the other clusters). Furthermore, the semantic proximity of cluster n°4 to the REFERENCE CLUSTER in TransDutch$_{ENG}$ is confirmed by the similar distance of the lexemes in both clusters to the abstract prototypes of other clusters. For TransDutch$_{FR}$, the semantic proximity between the cluster containing SPECIFIC ACTION and ACTION to the REFERENCE CLUSTER is also confirmed by the equal distances of the lexemes of both clusters to the abstract prototypes of the other clusters. In all three visualizations, nouns and verbs are clustered separately. However, we cannot maintain that clustering is totally independent of word class, since in the TransDutch fields, *eerst* becomes part of the REFERENCE CLUSTER and thus clusters with lexemes of a distinct word class.

4.5 Levelling out

The previous section has provided us with a number of insights with respect to the prototype-based organization of the clusters and the lexemes in each of the fields of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$ on the basis of *centroids* and *medoids*. These insights will now be used to see whether translation has impacted the organization of the fields on the semasiological or the onomasiological level and whether or not *levelling out* has taken place.

On the semasiological level, we will assess the changes in the distances of the clusters' *centroids* to the zero-point of the semantic space (considered as the prototypical center) they belong to amongst the different varieties. If the prototype-based organization of those meanings in translated Dutch differs from that in non-translated Dutch, and if this difference furthermore consists in *beginnen* having fewer different meaning differentiations in translated language compared to *beginnen* in non-translated Dutch, we can call the phenomenon semasiological *levelling out*.

On the onomasiological level, we will assess the changes in the distances of the lexemes in each cluster to the *centroid* (the abstract prototype) of the cluster they belong to. We will investigate whether the prototype-based organization of the lexemes in each cluster (with each cluster expressing a particular meaning differentiation) in translated Dutch differs from that in non-translated Dutch. Our method does however not allow us to investigate whether a given concept is expressed by *fewer* lexemes in translated Dutch compared to the same concept in non-translated Dutch, because the total number of lexemes within each semantic field is kept stable over all visualizations (see section 3.4.3)[57]. Observations on the onomasiological level will inform us about differences in the prototype-based organization of each cluster and possible changes in near-synonymy relationships between the lexemes in the semantic field under the influence of translation.

We first give a schematic overview of our observations on both the semasiological and the onomasiological level. The changes between the field of SourceDutch on the one hand and the fields of TransDutch$_{ENG}$ and TransDutch$_{FR}$ will be described subsequently.

[57] Since the number of lexemes is kept stable, any concept expressed by fewer lexemes would necessarily lead to another concept being expressed by more lexemes.

SourceDutch

Cluster n°	Meta-label(s)	Lexemes in cluster	Semasiological phenomena	Onomasiological phenomena
3	REFERENCE CLUSTER / GENERAL ONSET	*opstarten, starten, van start gaan, beginnen, gaan*	• closest to prototypical center ACTION STATE AFTER ONSET	• competition between *beginnen* and *starten* for position closest to the abstract prototype
2	GENERAL ONSET (NOUN)	*start, aanvang, begin*	• second closest to prototypical center • closest to REFERENCE CLUSTER	• competition between *start* and *begin* for position closest to the abstract prototype
1	SPECIFIC ACTION	*oprichten, opzetten*		
5	NON-LEXICALIZED INCHOATIVITY	*komen, krijgen, worden*		
4	ONSET OF ABSTRACT PROCESSES	*ontstaan, openen*		
6		*eerst*		

TransDutch$_{ENG}$

Cluster n°	**Meta-label(s)**	**Lexemes in cluster**	**Semasiological phenomena and changes**	**Onomasiological phenomena and changes**
3	REFERENCE CLUSTER / GENERAL ONSET	*eerst, van start gaan, beginnen, krijgen, starten, gaan, worden*	• closest to prototypical center • + eerst • + NON-LEXICALIZED INCHOATIVITY ACTION vs. STATE AFTER ONSET unclear	• *beginnen* closest to abstract prototype (< SourceDutch < TransDutch$_{FR}$) • more lexemes (<-> SourceDutch) • distance to abstract prototype: beginnen < gaan < krijgen < worden < starten < van start gaan
4	NO LABEL	*komen, opstarten*	• second closest to prototypical center	
2	ONSET (NOUN)	*begin*		• *begin* closest to abstract prototype (< SourceDutch < TransDutch$_{FR}$)
6	ONSET (NOUN)	*aanvang, start*	• closer to prototypical center than cluster n°2	• larger difference in distance to abstract prototype between *aanvang* and *start* (<-> SourceDutch)
1	SPECIFIC ACTION	*oprichten, opzetten*		• larger difference in distance to abstract prototype between *oprichten* and *opzetten* (<-> SourceDutch)
5	ONSET OF ABSTRACT PROCESSES	*ontstaan, openen*		• smaller difference in distance to abstract prototype between *openen* and *ontstaan* (<-> SourceDutch)

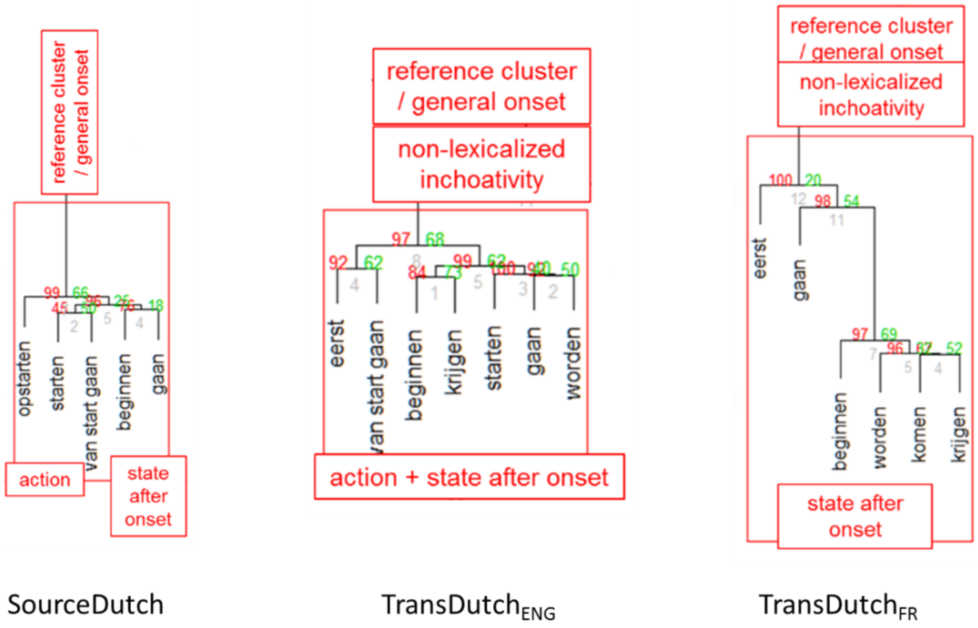
TransDutch$_{\text{FR}}$

Cluster n°	Meta-label(s)	Lexemes in cluster	Semasiological phenomena and changes	Onomasiological phenomena and changes
4	REFERENCE CLUSTER	*eerst, gaan, beginnen, worden, komen, krijgen*	• closest to prototypical center • + eerst • + NON-LEXICALIZED INCHOATIVITY STATE AFTER ONSET	• *beginnen* furthest away from abstract prototype (> SourceDutch > TransDutch$_{\text{ENG}}$) • more lexemes <-> SourceDutch
3	SPECIFIC ACTION	*opzetten,* *oprichten,* *opstarten,* *starten,* *van start gaan*	• second closest to prototypical center ACTION	• + *opstarten* • larger difference in distance to abstract prototype between *oprichten* and *opzetten* (<-> SourceDutch)
1	ONSET (NOUN)	*begin, aanvang, gaan*		• distance to prototype: *begin* < *aanvang* < *start*
2	ONSET OF ABSTRACT PROCESSES	*ontstaan, openen*		• smaller difference in distance to prototype between *openen* and *ontstaan* (<-> SourceDutch) • distance to prototype: *ontstaan* < *openen*

4.5.1 Semasiological levelling out

On the semasiological level, we observe the following changes:

- REFERENCE CLUSTER/GENERAL ONSET (Figure 84):
  - In TransDutch$_{ENG}$, the REFERENCE CLUSTER contains the meaning distinctions *eerst* and NON-LEXICALIZED INCHOATIVITY in addition to GENERAL ONSET (the only meta-label for this cluster in SourceDutch). The distinction between ACTION and STATE AFTER ONSET remains unclear on the semasiological level for TransDutch$_{ENG}$.
  - In TransDutch$_{FR}$, the REFERENCE CLUSTER contains the meaning distinctions *eerst* and NON-LEXICALIZED INCHOATIVITY in addition to GENERAL ONSET (the only meta-label for this cluster in SourceDutch). It does not, however, contain the meaning distinction ACTION.
  - In both TransDutch visualizations, more meaning distinctions become part of the REFERENCE CLUSTER compared to SourceDutch. In both TransDutch fields, the meaning distinctions *eerst* and NON-LEXICALIZED INCHOATIVITY become part of the REFERENCE CLUSTER, implying that they are used more prominently in TransDutch compared to SourceDutch.

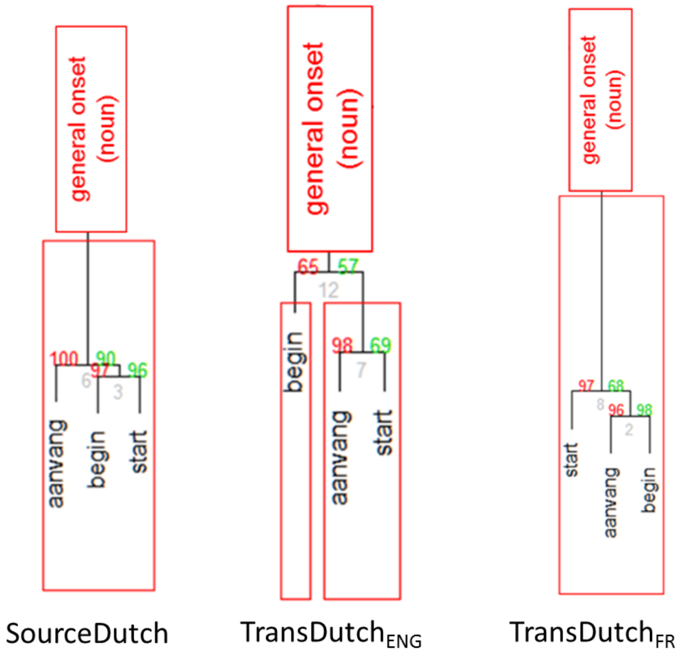

reference cluster / general onset
opstarten
starten
van start gaan
beginnen
gaan
action
state after onset
reference cluster / general onset
non-lexicalized inchoativity
eerst
van start gaan
beginnen
krijgen
starten
gaan
worden
action + state after onset
reference cluster / general onset
non-lexicalized inchoativity
eerst
gaan
beginnen
worden
komen
krijgen
state after onset

SourceDutch TransDutch$_{ENG}$ TransDutch$_{FR}$

Figure 84 REFERENCE CLUSTER/GENERAL ONSET of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$

- GENERAL ONSET (NOUN) (Figure 85)

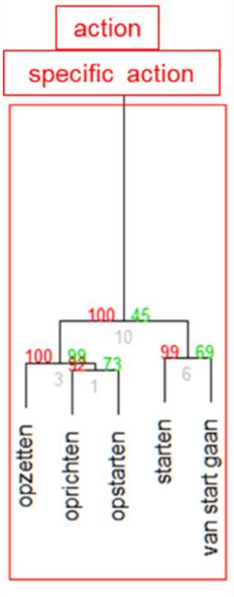
    - In TransDutch$_{ENG}$, *begin* forms a distinct cluster, whereas in SourceDutch, *begin* was part of GENERAL ONSET (NOUN). This division on the semasiological level suggests an additional meaning distinction within GENERAL ONSET (NOUN) in TransDutch$_{ENG}$.



SourceDutch TransDutch$_{ENG}$ TransDutch$_{FR}$

Figure 85 GENERAL ONSET (noun) of SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$

- ACTION (Figure 86)
    - In TransDutch$_{FR}$, a cluster is formed containing ACTION and SPECIFIC ACTION. This new cluster (meaning distinction) emphasizes the dynamic nature (the common denominator of ACTION and SPECIFIC ACTION) of the verbs it contains. In addition, the distinction between ACTION and STATE AFTER ONSET becomes more clearly marked in TransDutch$_{FR}$, compared to both SourceDutch and TransDutch$_{ENG}$ since ACTION and STATE AFTER ONSET now pertain to separate clusters.



TransDutch$_{FR}$

Figure 86 ACTION/SPECIFIC ACTION for TransDutch$_{FR}$

If we take a semasiological outlook, we can conclude that in translation, the meaning distinctions revealed by the different clusters do indeed differ from those in SourceDutch. In both TransDutch fields, some of the meaning distinctions that had been discerned for SourceDutch are now conflated in the REFERENCE CLUSTER. The cluster of GENERAL ONSET in both TransDutch fields thus 'absorbes' a certain amount of the semasiological variation that was present in SourceDutch. If we take the meanings distinguished on the basis of SourceDutch as a point of reference, we see that fewer of those meanings are also distinguished by the TransDutch fields. As a consequence, we could speak of a presence of semantic *levelling out* on the semasiological level. Two observations seem to go against this statement. First, for TransDutch$_{FR}$, we observe that on the one hand, the meaning distinction between ACTION and STATE AFTER ONSET is emphasized compared to SourceDutch (ACTION and STATE AFTER ONSET are now part of two distinct clusters, implying no *levelling out*). On the other hand, the conflation of ACTION and SPECIFIC ACTION erases the meaning distinction between ACTION and SPECIFIC ACTION, so that *levelling out* on the semasiological level can be claimed. Second, in TransDutch$_{ENG}$, a meaning distinction containing only *begin* is suggested, and a second one containing *opstarten* and *komen* is also discerned, implying more semasiological specification than in SourceDutch.

4.5.2 Onomasiological changes in the prototype-based organization

On the onomasiological level, we observe the following changes:

- REFERENCE CLUSTER/GENERAL ONSET:

- The unclear distinction between ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$ is clarified on the onomasiological level: the distances of the lexemes to the abstract prototype (*centroid*) of the REFERENCE CLUSTER of TransDutch$_{ENG}$ show that STATE AFTER ONSET verbs (*beginnen* and *gaan*) are closer to the abstract prototype, but that ACTION verbs (*starten* and *van start gaan*) are situated much further away from the abstract prototype. This organization is different from SourceDutch, where *beginnen* and *starten* are both on a minimal distance to the abstract prototype. In other words, the difference in distance to the prototype between *starten* and *beginnen* becomes larger in TransDutch$_{ENG}$, compared to SourceDutch. In TransDutch$_{FR}$, *starten* and *beginnen* are part of different clusters (and hence more dissimilar). We can conclude that for both TransDutch semantic representations, *beginnen* and *starten* are less near-synonymous than in SourceDutch.

- GENERAL ONSET (NOUN)*:*
    - Similar to the situation for *beginnen* and *starten*, we see that for *start* and *begin* a competition for the position closest to the abstract prototype is going on in SourceDutch. In both TransDutch fields, the competition between *begin* and *start* is less present: in TransDutch$_{ENG}$, a separate cluster with *begin* appears, and in TransDutch$_{FR}$, we see that *begin* is closest to the abstract prototype, but that *start* is situated much further away from the abstract prototype. We can conclude that *begin* and *start* are less near-synonymous in both TransDutch fields compared to SourceDutch.
- SPECIFIC ACTION:
    - A competition for the position closest to the abstract prototype is also going on between *oprichten* and *opzetten*. A similar situation appears here: in SourceDutch, both lexemes are extremely close to the abstract prototype, whereas in the TransDutch fields, the difference in distance to the abstract prototype increases, implying that the lexemes are less near-synonymous in TransDutch compared to SourceDutch.

From an onomasiological point of view, we observe small differences in the prototype-based organization of the lexemes in TransDutch compared to SourceDutch. We see that *starten* and *beginnen* become less near-synonymous (the difference in distance between the lexemes with respect to the prototype becomes larger) in both TransDutch fields. The same observation can be made for *start* and *begin*: the two lexemes are more near-synonymous in SourceDutch, but less near-synonymous in TransDutch$_{ENG}$ and TransDutch$_{FR}$. This is also observed for *oprichten* and *opzetten*: they are more synonymous in SourceDutch compared to TransDutch$_{ENG}$ and TransDutch$_{FR}$. Although the joint clustering of (pairs of) lexemes of course confirms the synonymy between the lexemes, we could conclude that lexemes which are near-synonyms in SourceDutch (such as *starten* and *beginnen*, *start* and *begin*, *oprichten* and *opzetten*) tend to become less near-synonymous in translated language. Note that we only observe this trend for lexemes which are near-synonyms in SourceDutch (both very close to the abstract prototype). For lexemes pertaining to the same cluster (which can also be considered as synonyms given their joint clustering) which show larger differences in distance to the prototype in SourceDutch (indicating less near-synonymy) such as *ontstaan* and *openen*, the difference in distance to the abstract prototype is not increased by translation.

4.6 Shining through

4.6.1 Semasiological shining through

We investigate semasiological *shining through* (source language influence on the meaning distinctions in translated language) by comparing the meaning distinctions in translated language to those present in the source language of the translation. We therefore visualize the semantic fields of the closest equivalents of *beginnen* in the source languages of TransDutch$_{ENG}$ and TransDutch$_{FR}$: SourceEnglish *to begin* and SourceFrench *commencer*.

Ideally, we should first provide an analysis of SourceEnglish and SourceFrench following the exact same steps as for SourceDutch (a statistical visualization, followed by a description of the prototype-based organization of the semantic field on both the semasiological and the onomasiological level, leading to an in depth description and interpretation of the semantic field) before we compare the different meaning distinctions (clusters) in the fields of *to begin* and *commencer* to the meaning distinctions in TransDutch$_{ENG}$ and TransDutch$_{FR}$. In this section, we will however only present the visual output of the HAC (carried out on the output of a CA, according to the exact same procedure as described in chapter 3) for SourceEnglish and SourceFrench without providing a lengthy discussion of the prototype-based organization of those two fields. A full description – the ideal scenario – would require a complete contrastive comparison of the fields of SourceEnglish and SourceFrench (and SourceDutch) before the influence of SourceEnglish and SourceFrench on TransDutch$_{ENG}$ and TransDutch$_{FR}$ could be determined. Obviously, such a description would enhance our insights into the influence on the target language of attested differences between the source language and the target language semantic fields. Due to the contrastive scope of such a description as well as timely constraints, we will, however, present the visualizations of SourceEnglish and SourceFrench in the light of the possible explanations they could provide for a number of differences observed in the TransDutch$_{ENG}$ and TransDutch$_{FR}$ fields and which are possibly caused by specific source language influence.

4.6.1.1 Semasiological shining through of SourceEnglish

Three semasiological changes in TransDutch$_{ENG}$ (compared to SourceDutch) might have been influenced by existing meaning distinctions in SourceEnglish: (i) the separate clustering of *begin*, (ii) the separate clustering of *opstarten* and *komen*[58] and (iii) the unclear distinction between ACTION and STATE AFTER ONSET (on the semasiological level) in the REFERENCE CLUSTER of TransDutch$_{ENG}$. A source language influence could be claimed if, in SourceEnglish, a separate meaning distinction (cluster) containing the closest translational equivalent of *begin*, i.e. *beginning* was attested and/or a separate meaning distinction containing *to start up* and *to come* (the closest translational equivalents of *opstarten* and *komen*). If in SourceEnglish, the meaning distinction (possibly within the most central cluster of the analysis) between ACTION and STATE AFTER ONSET is equally unclear as in TransDutch$_{ENG}$, we could possibly interpret this as source language influence.

The semantic field of SourceEnglish is visualized on the basis of data retrieved via the SMM++ with *to begin* as initial lexeme and Dutch as a language B. We follow the exact same procedure as for SourceDutch. One important difference needs to be noted here: since the DPC does not contain data for the translation directions French to English and English to French, only one language can be used as a language B when an English initial lexeme is chosen, *in casu*, Dutch. The establishment of the data set for SourceEnglish is consequently only based on the *second T-image* of *to begin* with Dutch as a pivot language (recall that for SourceDutch, the data sets of the *second T-image* of beginnen$_{FR}$ and beginnen$_{ENG}$ were combined)[59]. The outcome of the SMM++ retrieval task rendered a set of 30 English lexemes (911 observations). We carry out a HAC on the output of the CA and choose a cluster solution with 5 clusters (*average silhouette width* 0.7).

[58] As we have seen that the clustering of *komen* is unstable (section 4.3.1), the analysis will mainly focus on *opstarten*.

[59] One could argue here that the semantic field of SourceEnglish is likely to be biased by the fact that the used data set is only based on the *second T-image* of *to begin* with Dutch as a source language. In order to solve this problem while maintaining our translational method, we would however need a tri-directional corpus (where all three languages can be used as languages B to carry out a SMM++) which we do not have at our disposal. Another solution would be to apply an alternative, distributional technique to visualize the SourceLanguage semantic fields which would only use monolingual (Dutch) data to create the data matrix (rather than the translations). A comparison of the translational and the distributional approach is provided in Vandevoorde et al. (2016) and shows that the patterns revealed by both methods are very similar.

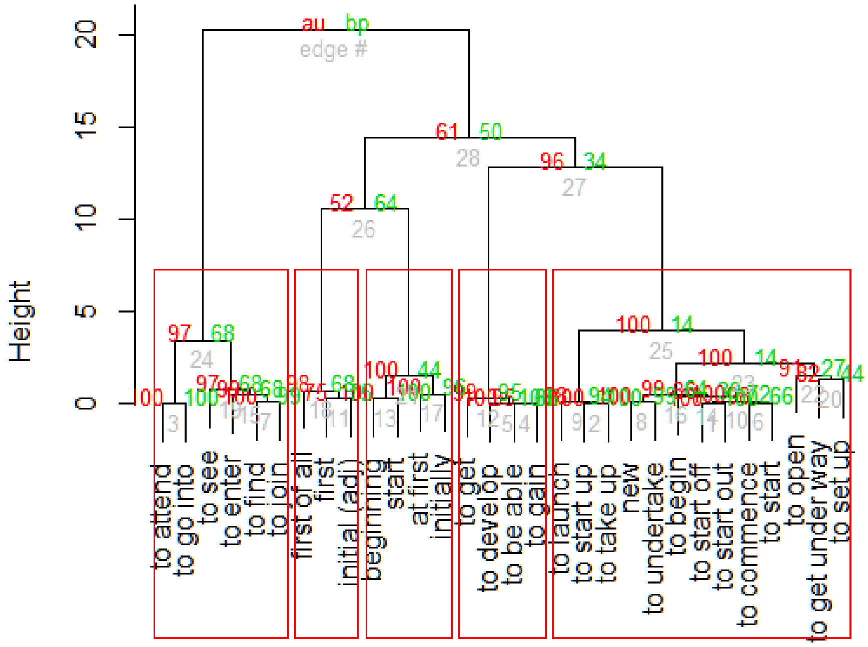


Figure 87 Dendrogram representing a semantic field of *to begin* for SourceEnglish

Looking at the dendrogram (Figure 87) for SourceEnglish, we see that *beginning* is part of a cluster with *start*, *at first* and *initially* so that no separate meaning distinction of *beginning* is implied by SourceEnglish. We can conclude that the separate meaning distinction of *begin* in TransDutch$_{\text{ENG}}$ is not triggered by an existing meaning distinction in SourceEnglish.

We also observe that *to start up* is part of the largest cluster (and most central one in the semantic space) of the analysis, containing both *to start* and *to begin*. The closest translational equivalent of *komen, to come* is not a lexeme in the SourceEnglish visualization[60]. We can again conclude that the separate meaning distinction of *komen* and *opstarten* in TransDutch$_{\text{ENG}}$ is not caused by an existing meaning distinction in SourceEnglish.

As for the unclear distinction between ACTION and STATE AFTER ONSET in TransDutch$_{\text{ENG}}$, we notice that in SourceEnglish, no clear division between ACTION and STATE AFTER ONSET is marked either. The prototypical ACTION verb *to start* and the prototypical STATE AFTER ONSET verb *to begin* are both part of the same, most central

[60] *Komen* is a verb which typically does not lexicalize inchoativity and draws its inchoative meaning from the context it is used in. As a consequence, its closest translational equivalent *to come* does not typically express inchoativity and is, unsurprisingly, not a member of the SourceEnglish field.

cluster (the outer right cluster in the dendrogram), although they belong to different sub-nodes (just as was the case for TransDutch$_{ENG}$ and SourceDutch). Semasiological *shining through* could be claimed here, although it must be admitted that - given the similar divide between ACTION and STATE AFTER ONSET in SourceDutch – the phenomenon could well be interpreted as semasiological *normalization* too (see section 4.7.1).

4.6.1.2 Semasiological shining through of SourceFrench

For TransDutch$_{FR,}$ we observed that the meaning distinctions ACTION and SPECIFIC ACTION are 'absorbed' by a new cluster. This new cluster emphasizes the (common) dynamic nature of the meaning distinctions it absorbed (while the specificity of the meaning distinctions indicated by ACTION and SPECIFIC ACTION is somewhat 'levelled out'). In addition, the distinction between ACTION and STATE AFTER ONSET is more emphasized in TransDutch$_{FR}$ (the labels are assigned to different clusters), compared to SourceDutch and SourceEnglish (where ACTION and STATE AFTER ONSET pertain to the REFERENCE CLUSTER). In this section, we now investigate whether source language influence has possibly caused this semasiological change.

The data for the visualization of SourceFrench were retrieved via the SMM++ with *commencer* as initial lexeme and Dutch as language B. Parallel to the field of SourceEnglish, the field of SourceFrench (Figure 88) is only based on data from the *second T-image* of *commencer* with Dutch as language B. The SMM++ retrieval task rendered a set of 25 French lexemes (824 observations). We carried out a HAC on the output of the CA. The chosen cluster solution with 4 clusters obtains an *average silhouette width* of 0.54.

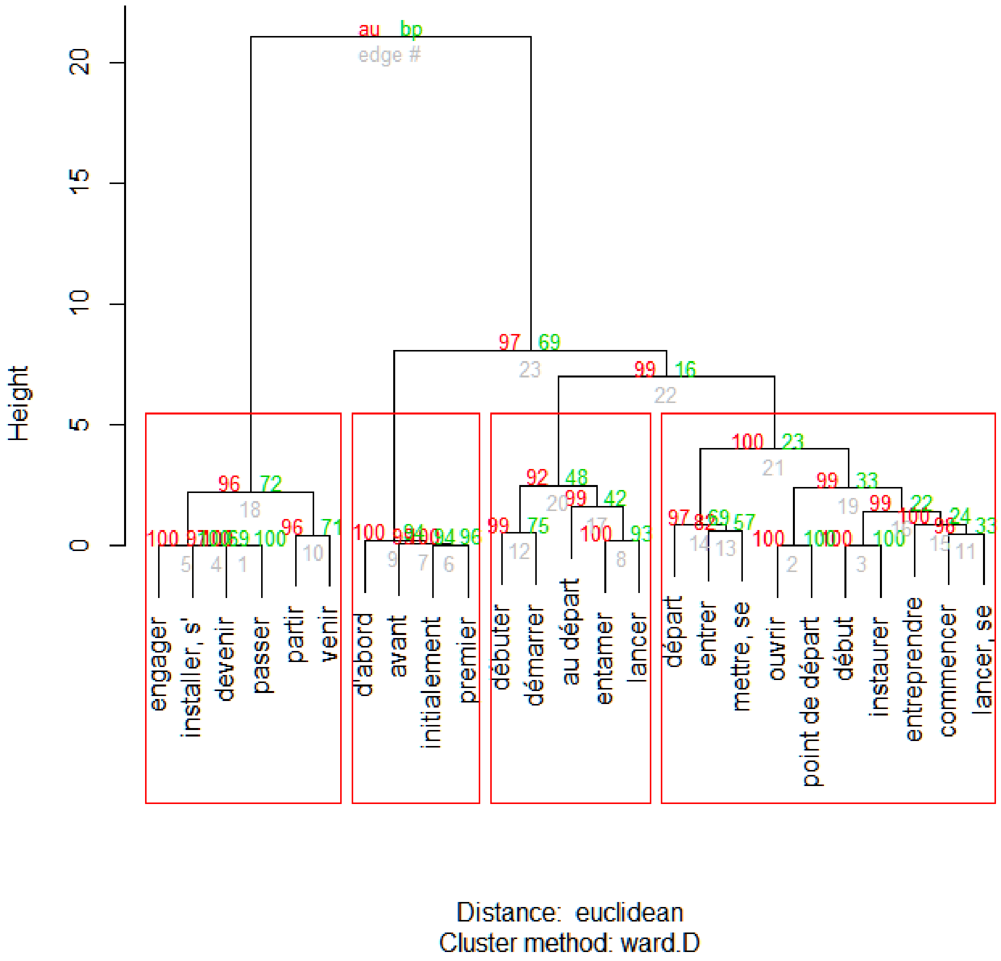


Figure 88 Dendrogram representing a semantic field of *commencer* for SourceFrench

Like in English (and Dutch), inchoativity in French is also thought to present the division between more dynamic ACTION verbs ("focusing on the transition from non-ACTION to ACTION") and more static STATE AFTER ONSET verbs ("indicating the start of a transformation") (Marque-Pucheu 1999, 241). Although Marque-Pucheu does not specify any particular verbs of inchoativity that are more typically used with the one rather than with the other verb type, clearly, *démarrer* [to start up], *entamer* [to start] and *débuter* [to begin, to start] are verbs that can be categorized as ACTION verbs (they are used with *moteur* [engine] for example), while *commencer* (the translational equivalent of *to begin*) seems to focus on the STATE AFTER ONSET. Within SourceFrench, we indeed observe a cluster containing these ACTION verbs *entamer*, *débuter*, *démarrer*, *au départ* [initially] and *lancer* [to launch]. If we now take a closer look at the cluster containing *commencer*, we see that some of the lexemes clustering with *commencer* indeed suggest that this cluster is focusing on the more static STATE AFTER ONSET. *Entrer*, for instance, can indicate *commencer à être dans un lieu, à un endroit, dans un état, dans une période* [to start being in a place, state, period...] (Grand Robert de la Langue Française, 2013), and *se mettre*, can mean *devenir quant à l'état psychique, la situation* [to become into a physical state, a situation] or – when followed by the preposition 'à' –

commencer à faire [to begin to do something]. It could be claimed that in SourceFrench, a clear meaning distinction is made between ACTION and STATE AFTER ONSET (they make up distinct clusters). The separate clustering of ACTION and STATE AFTER ONSET in TransDutch$_{FR}$ might then have been triggered by the distinct clustering of ACTION and STATE AFTER ONSET in SourceFrench as an instance of semasiological *shining through.*

However, in the same cluster of *commencer*, we find a number of lexemes which seem to be more related to business-like contexts (and could easily be labelled as SPECIFIC ACTION), such as *entreprendre* [to undertake] and *se lancer* [to launch oneself into]. These lexemes expressing SPECIFIC ACTION are clustering with STATE AFTER ONSET in SourceFrench, whereas in TransDutch$_{FR,}$ they form a cluster with ACTION. As a consequence, the joint clustering of ACTION and SPECIFIC ACTION cannot be explained on the basis of semasiological *shining through.*

As we announced in the introduction of this section, the provided interpretation is of course preliminary, and can only hint us towards possible instances of semasiological *shining through.* A more thorough analysis of the SourceEnglish and the SourceFrench field is needed to understand the mechanisms of source language influence on the TransDutch fields. For TransDutch$_{FR}$, for example, such an analysis would have to confirm or disaffirm whether the presumed distinction between ACTION and STATE AFTER ONSET does indeed correspond to the lexemes in the respective clusters of SourceFrench and / or whether the assumed joint clustering of SPECIFIC ACTION with STATE AFTER ONSET in SourceFrench can indeed be claimed.

4.6.2 Onomasiological shining through

In this section, we present two additional visualizations for TransDutch$_{ENG}$ and TransDutch$_{FR}$ which contain the English and French source language lexemes together with the Dutch target language lexemes. In this way, we can see whether onomasiological *shining through* is taking place (whether the organization of the lexical items in the meaning distinctions in the fields of TransDutch$_{ENG}$ and TransDutch$_{FR}$ is influenced by a specific underlying source language lexeme). Rather than describing the influence of each underlying English or French source language lexeme, we will focus on those instances where a specific source language lexeme might explain a change in the organization of the lexemes in TransDutch$_{ENG}$ or TransDutch$_{FR}$ compared to SourceDutch.

4.6.2.1 Onomasiological shining through of English

In section 4.6.1, we saw that semasiological *shining through* could not account for the separate clustering of *begin*, nor for the separate clustering of *opstarten* in TransDutch$_{ENG}$.

We now explore whether this separate clustering could be the result of an instance of onomasiological *shining through* (the specific influence of a source language lexeme).

The simultaneous visualization of the source and target language lexemes in a single space is carried out via a Multiple Correspondence Analysis on a Burt table (Greenacre 2006, 2007) (see section 3.6.4). We use the output of the Multiple Correspondence Analysis, as the input for a HAC. Although the visualization of the HAC on the output of a MCA at first sight looks quite different from the dendrogram representing a semantic field of *beginnen* for TransDutch$_{ENG}$, the two visualizations do depict the same reality: the clustering of the Dutch lexemes in Figure 89[61] below is identical to that in Figure 64 (all clusters correspond to either a cluster or a sub-node)[62].

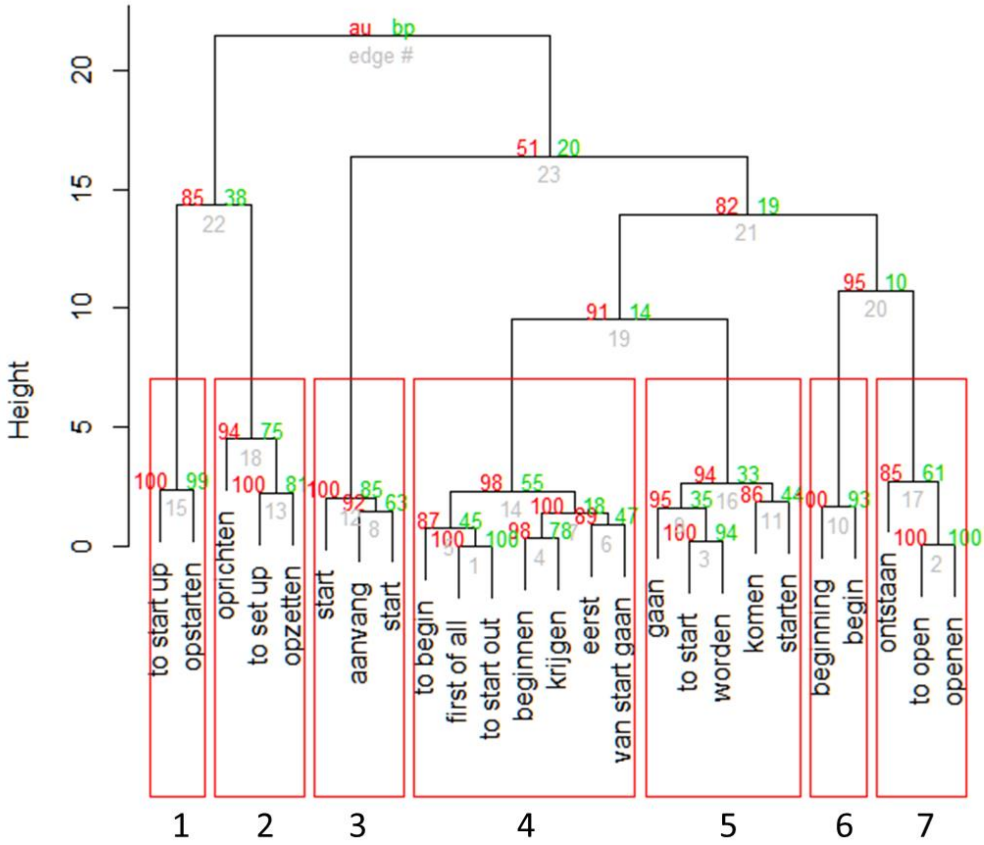


Figure 89 Representation of HAC on the MCA for TransDutch$_{ENG}$

[61] The clusters are numbered from left to right.

[62] Note that the lexemes from cluster n°4 from TransDutch$_{ENG}$ (*komen* and *opstarten*) are now spread over two different clusters – this was to be expected given the 'unstable' clustering in TransDutch$_{ENG}$ of those two lexemes. The lexemes of the REFERENCE CLUSTER of TransDutch$_{ENG}$ (cluster n°3) are now spread over two clusters, which are joined in a higher, slightly less significant node within this visualization.

On a general level, it immediately strikes us that all English source language lexemes are clustered together with their Dutch close cognate whenever the latter is present in the analysis (only *first of all* and *to start out* do not have direct close cognate amongst the Dutch lexemes). We discern the following pairs: *beginning-begin*; *start-start*; *to open-openen*; *to begin-beginnen*; *to start-starten*; *to start up-opstarten*; *to set up-opzetten*.

With regard to Dutch *begin*, we see that the lexeme is clustered with its English close cognate *beginning* (cluster n°6), revealing the preference of *begin* to apply as a translation of *beginning*. The same goes for *opstarten*, which is clustered here with its close cognate *to start up*. In both cases, the underlying English source language lexemes seem to trigger the separate clustering of *begin* and *opstarten*. In this way, an influence on the onomasiological level seems to provoke semasiological change in TransDutch$_{ENG}$ compared to SourceDutch. This onomasiological *shining through* is very likely to be triggered by the strong semantic relatedness between the elements of pairs of close cognates such as *begin – beginning* and *opstarten – to start up*.

4.6.2.2 Onomasiological shining through of French

In section 4.6.1.2, we tentatively accounted for the clear (over-emphasized with respect to SourceDutch) meaning distinction between ACTION and STATE AFTER ONSET in TransDutch$_{FR}$ via semasiological *shining through*. The joint clustering of ACTION and SPECIFIC ACTION could however not be explained on the semasiological level. In this section, we want to investigate whether the joint clustering of ACTION and SPECIFIC ACTION could be the result of an instance of onomasiological *shining through* (the influence of a specific source language lexeme on the organization of the lexemes within a cluster / meaning distinction).

The clustering of the Dutch lexemes presented in the visualization in Figure 90[63] shows the same semantic field of *beginnen* for TransDutch$_{FR}$ as the dendrogram of the HAC for TransDutch$_{FR}$ in Figure 76 (all clusters correspond to either a cluster or a sub-node).

[63] The clusters are numbered from left to right.

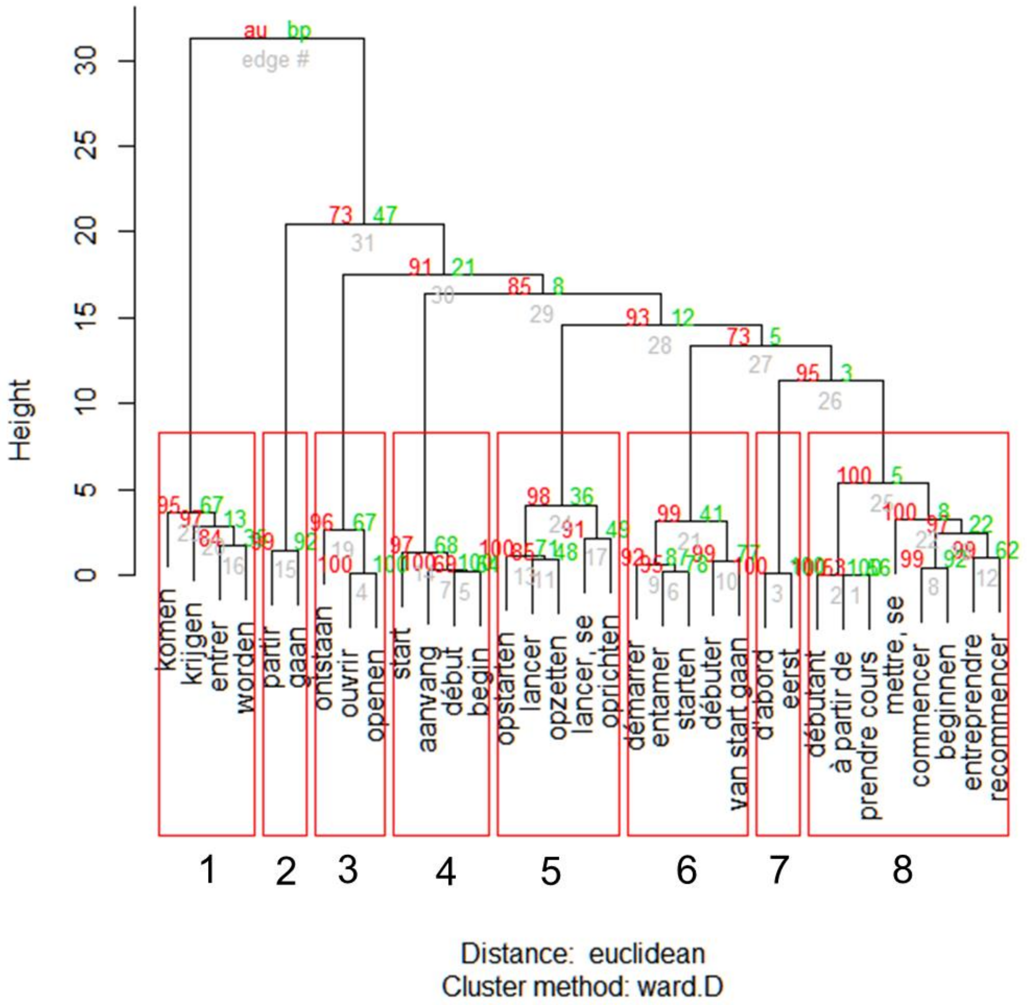


Figure 90 Representation of HAC on the MCA for TransDutch$_{FR}$

The cluster reuniting SPECIFIC ACTION and ACTION in the HAC visualization in Figure 76 corresponds to clusters n°5 and 6 in Figure 90. We see that the Dutch lexemes *opstarten*, *oprichten* and *opzetten* in cluster n°5 (SPECIFIC ACTION) are often translations of *lancer* [to launch] and *se lancer* [to launch, to go into]. The Dutch lexemes *starten* and *van start gaan* in cluster n°6 (ACTION) are often translations of *entamer*, *démarrer* and *débuter*. This analysis shows that specific source language lexemes are underlying either the meaning distinction ACTION or SPECIFIC ACTION. A distinct clustering of ACTION and SPECIFIC ACTION in TransDutch$_{FR}$ would be expected on the basis of this information. The fact that this is not the case (and that ACTION and SPECIFIC ACTION cluster together in TransDutch$_{FR}$) argues against onomasiological *shining through.*

If we now reconnect the information gathered on the onomasiological level to the semasiological level, we can gain some additional insights. The French source language lexemes in cluster n°6 correspond to the ones pertaining to the cluster ACTION in SourceFrench. However, the underlying lexemes of the cluster of SPECIFIC ACTION (n°5) in the above analysis (*lancer* and *se lancer*) did not form a distinct cluster in SourceFrench (*lancer* was part of the ACTION cluster and *se lancer* was part of the STATE

AFTER ONSET cluster). This could mean that no meaning distinction for SPECIFIC ACTION is discerned in SourceFrench (the lexemes expressing SPECIFIC ACTION are part of different clusters) and in turn explain – as semasiological *shining through* – why in TransDutch$_{FR}$, SPECIFIC ACTION is no longer forming a separate cluster.

Although we cannot be sure about how exactly the clustering of ACTION with SPECIFIC ACTION has come about in TransDutch$_{FR}$, we can, however, be quite sure that it is triggered by a change on the semasiological level, possibly by semasiological *levelling out*. It once again becomes clear, however, that the provided interpretations are tentative and can by no means detect with certainty instances of *shining through*.

4.7 Normalization

4.7.1 Semasiological normalization

We will now investigate semasiological *normalization* (target language influence on the meaning distinctions in translated language) by comparing the meaning distinctions present in the visualizations of SourceDutch to the meaning distinctions in TransDutch$_{ENG}$ and TransDutch$_{FR}$. If a same meaning distinction appears in TransDutch$_{ENG}$ and TransDutch$_{FR}$ and this organization is in addition similar or identical to the organization in SourceDutch, there is a fair chance that the TransDutch fields are 'conforming' to the SourceDutch field, yielding evidence for semasiological *normalization*.

For the semantic field of inchoativity, one clear example of semasiological *normalization* is the cluster ONSET OF ABSTRACT PROCESSES. This meaning distinction is present in both TransDutch visualizations and an identical cluster can be found in SourceDutch.

A second, possible instance of semasiological *normalization* concerns the meaning distinction between ACTION and STATE AFTER ONSET within the REFERENCE CLUSTER of TransDutch$_{ENG}$. Although we showed that this could be interpreted as semasiological *shining through* (see section 4.6.1.1), semasiological *normalization* could also be claimed here since in SourceDutch, ACTION and STATE AFTER ONSET also pertain to the REFERENCE CLUSTER. The same now holds for the separate clustering of ACTION and STATE AFTER ONSET in TransDutch$_{FR}$: it could equally be interpreted as an (over-)*normalization* of the distinction in SourceDutch. The fact that a same phenomenon can be interpreted as either semasiological *normalization* or semasiological *shining through* should not worry us. In fact, it confirms that translated language comes into being within some kind of 'continuum', of which the one end is over-*normalization* and the other end *shining through* (Hansen-Schirra & Steiner 2012, 272). Phenomena which are situated in the center of this continuum (of which the case of ACTION – STATE AFTER ONSET might be a good example) can consequently be interpreted as either *shining through* or *normalization.*

4.7.2 Onomasiological normalization

Onomasiological *normalization* (target language influence on the prototype-based organization of the lexemes within each meaning distinction of *beginnen*) will be investigated by comparing the prototype-based organization of the lexemes in each meaning distinction in SourceDutch to the organization of the lexemes in each meaning distinction in TransDutch$_{ENG}$ and TransDutch$_{FR.}$ If the same organization of lexemes appears in TransDutch$_{ENG}$ and TransDutch$_{FR}$ and this organization is in addition similar or identical to the organization in SourceDutch, there is a fair chance that the TransDutch fields are 'conforming' to the SourceDutch field, yielding evidence for onomasiological *normalization.*

The presence of onomasiological *normalization* can only be investigated for clusters which contain the same lexemes in a TransDutch field and SourceDutch. Onomasiological *normalization* cannot be determined between clusters that are not identical since the addition or removal of one or more lexemes will as such already influence the prototype-based organization of the lexemes within this cluster (and the possible influence of the target language on the structure cannot be teased apart any longer).

Both in TransDutch$_{ENG}$ and in SourceDutch, we discern the cluster SPECIFIC ACTION, containing the lexemes *oprichten* and *opzetten.* As such, the joint clustering of the lexemes in both varieties confirms the synonymy between the lexemes in both fields. If we look at the distance to the prototype in either variety, we see that in SourceDutch, both lexemes are very close to the abstract prototype (the *centroid*) of the cluster they belong to (*opzetten* is at `0.06749455` of the *centroid* in SourceDutch and at `0.2476172` for TransDutch$_{ENG}$, *oprichten* is at `0.02952887` of the *centroid* in SourceDutch and at `0.2004520` in TransDutch$_{ENG}$). Although the difference in distance to the prototype between *oprichten* and *opzetten* increases slightly in TransDutch$_{ENG}$ (they are slightly less near-synonymous in TransDutch$_{ENG}$) we could claim a case of onomasiological *normalization* here (the prototype-based organization of the lexemes within TransDutch$_{ENG}$ is conforming to SourceDutch).

In SourceDutch and TransDutch$_{FR}$, we see that the cluster GENERAL ONSET (NOUN) contains the lexemes *begin*, *start* and *aanvang.* Again, the identical clustering already confirms their near-synonymy in both fields. In SourceDutch, *begin* (`0.08908944`) is the closest lexeme to the abstract prototype, *start* (`0.20740218`) is situated slightly further away and *aanvang* (0.55330205) still somewhat further away. In TransDutch$_{FR}$, we see that *begin* (0.05884857) is the closest lexeme to the abstract prototype, but we see that *aanvang* (0.12955053) is now much closer to the abstract prototype than *start* (0.54160901) is. For this case, no onomasiological *normalization* can be claimed since the prototype-based organization of the lexemes in TransDutch$_{FR}$ does not conform to that in SourceDutch.

Finally, in all three fields, we discern an identical cluster containing the lexemes *ontstaan* and *openen.* If we assess the difference in distance to the prototype, we see that in SourceDutch, *openen* is very close to the abstract prototype (0.1718314), and *ontstaan* is situated much further away (1.3471583). These lexemes are then less near-synonymous than *oprichten* and *opzetten* for example (which are both at a minimal distance of their abstract prototype). For TransDutch$_{ENG}$, we now see that the difference in distance to the abstract prototype slightly decreases (we could say that *openen* (0.1067802) and *ontstaan* (1.1745826) become slightly more synonymous in TransDutch$_{ENG}$). This could consequently be interpreted as an instance of *normalization*: the prototype-based organization of the lexemes in this cluster in TransDutch$_{ENG}$ is conforming (and even slightly 'exaggerating') the prototype-based structure of the lexemes in the same cluster in SourceDutch. For TransDutch$_{FR}$, however, we see that *ontstaan* (0.0593305) is now the closest lexeme to the prototype, and that *openen* (0.9492880) is situated further away from the abstract prototype. This argues against onomasiological *normalization.*

4.8 Conclusion

In this chapter, we have provided a detailed interpretation of the visualizations of the semantic field of *beginnen* / inchoativity for SourceDutch, TransDutch$_{ENG}$ and TransDutch$_{FR}$. On the basis of these interpretations we have explored whether a number of universal tendencies of translation also hold on the semantic level.

We can conclude that the meanings expressed by *beginnen* do differ in translated language compared to non-translated language and hence that there are differences between the fields of translated and non-translated Dutch inchoativity on the semasiological level. We also observed that the prototype-based organization of lexemes within the different meaning distinctions differed in translated language, compared to non-translated language, so differences are also attested on the onomasiological level.

We have found evidence for semantic *levelling out* on the semasiological level in translated Dutch. In both TransDutch fields, some of the semasiological variation present in SourceDutch was 'absorbed' by the REFERENCE CLUSTER. On the onomasiological level, we concluded that a number of near-synonymous pairs in SourceDutch seemed to become somewhat less near-synonymous in translated Dutch.

The joint clustering of ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$ and the separate clustering of ACTION and STATE AFTER ONSET in TransDutch$_{FR}$ could be explained as *shining through* on the semasiological level. For TransDutch$_{FR}$, the joint clustering of ACTION and SPECIFIC ACTION could also be interpreted as semasiological *shining through*. The separate clustering of *begin* and *opstarten* in TransDutch$_{ENG}$ could be explained as onomasiological *shining through.*

We detected semasiological *normalization* for the cluster ONSET OF ABSTRACT PROCESSES. We further noted that the specific clustering of ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$ (in the REFERENCE CLUSTER) and in TransDutch$_{FR}$ (in separate clusters) could also be explained as different degrees of semasiological *normalization.* Finally, the lexemes *oprichten* and *opzetten* show onomasiological *normalization* in TransDutch$_{ENG}$.

Different (and sometimes seemingly contradictory) tendencies are thus at play here and seem to determine the structure of the semantic fields: larger tendencies of *levelling out* on the semasiological level as well as *shining-through* seem to act upon the TransDutch fields. This chapter has provided a number of insights with respect to the possible influence of *levelling out, normalization* and *shining through* on both the semasiological and the onomasiological level. It does, however, not explain why for some phenomena, *levelling out* on the semasiological level seems to prevail and for others, onomasiological *shining through* seems to be determinant for the clustering. In the next chapter, we will try to understand how such seemingly contradictory mechanisms can act upon a same semantic representation. We will do so by interpreting our results within more broad, cognitive-translational, explanatory frameworks from cognitive translation studies and bilingualism.

Chapter 5
Cognitive Explorations

5.1 Introduction

In the previous chapter, we have shown how the established method can be used to create visualizations of semantic fields of translated and non-translated language which can consequently be compared to each other. The observed differences between the translated and non-translated semantic fields of inchoativity were tentatively explained by applying the framework of *translation universals* on the semantic level. Although the observations could be 'fitted into' the translation universal framework, this does not as such explain *why* these – sometimes surprising and seemingly contradictory – phenomena appear (the observed phenomena can be connected to universal tendencies of translation, but the fact that an observed phenomenon can be understood as a universal tendency does not explain why it appears in the first place nor where it comes from). In this chapter, we will therefore look for cognitive explanations for our main observations described in chapter 4: (i) the overall *levelling out* on the semasiological level in translated language; (ii) the instances of onomasiological *shining through* in TransDutch$_{\text{ENG}}$ (the separate clustering of *begin* and *opstarten*); (iii) the semasiological *shining through* or *normalization* causing the joint clustering (in TransDutch$_{\text{ENG}}$) or separate clustering (in TransDutch$_{\text{FR}}$) of ACTION and STATE AFTER ONSET and (iv) the joint clustering of ACTION and SPECIFIC ACTION in TransDutch$_{\text{FR}}$ under influence of semasiological *shining through*.

In this chapter, we will put forward two models that can possibly provide us with cognitive explanations for these findings. First, we will try to understand our results in the light of Halverson's (2003, 2010, 2013, forthcoming) Gravitational Pull Hypothesis (section 5.2) (hence: GPH). In the subsequent section 5.3, we will try to interpret our results a second time, now on the basis of a cognitive-explanatory model from neurolinguistics (Paradis 2004, 2007) which was introduced in TS by Juliane House (2013). These models are two of the few that have been put to the fore within cognitive translation studies. However, to date, few attempts have been made to apply them as explanatory frameworks for empirical studies in TS. Before we try to account for our results using either framework, we will, in the remainder of this section, zoom in on how cognitive explanations can be linked to corpus data (section 5.1.1), and more specifically to semantic fields (section 5.1.2). In section 5.1.3, we will compare the starting points of the two models before we present and apply them to our results in sections 5.2 and 5.3.

5.1.1 Linking cognitive explanations to corpus data

Before we venture into this search for cognitive explanations, we first need to clarify how evidence from corpus data can be linked to cognitive *explanations*. In chapter 2, we have substantiated our choice to connect a corpus linguistic methodology with a cognitive linguistic theoretical framework. We have equally discussed how the re-integration of the study of meaning within Translation Studies was only possible within the so-called cognitive turn in TS. More particularly, a linguistic-cognitive outlook seemed a much needed basis for “a theoretically based description and explanation of how strategies of comprehending, problem solving and decision making with reference to the texts that translators handle come about in their bilingual minds” (House 2013, 48). In the previous chapters the focus has been on the first aspect quoted by House, a theoretically based description. In the current chapter, our aim is to put forward theoretically based cognitive explanations for the results obtained within this corpus-based cognitive study of translation.

Cognitive explanations “emphasize that the usage of a given form is governed by principles that ensure ease of production and processing” (Arppe et al. 2010, 20). Off-line linguistic data are not normally expected to provide evidence for such kinds of principles. Arppe and colleagues claim, however, that evidence from experimental research would not necessarily serve this goal better. They point out that diverging evidence from corpus data and experimental research does not automatically dismiss the corpus evidence. Giving an example of the link between ease of activation and diverging corpus and experimental results, they conclude that:

> [t]he fact that the most frequent corpus sense in the study [...] was not among the first that came to mind in the sentence production experiment may just as well reflect a limitation of the experimental design rather than prove that frequency does not determine ease of activation [...] when subjects are led to think about word meanings, it is perhaps not surprising that the most frequent responses do

> not involve semantically light to near-empty senses of the prime (Arppe et al. 2010, 11-12).

Elicitation protocols are thus not thought to “provide an *a priori* more reliable probe into cognitive processes than other methods” (Arppe et al. 2010, 12). Arguably, converging evidence from different types of research will enable the researcher to make a stronger plea in favor of the advanced hypothesis, but diverging evidence does not automatically disprove the corpus evidence. Hence, the link between corpus results and cognitive explanations is not necessarily less plausible than the link between results of experimental research and such explanations.

In both cases, caution is recommended as to how one links the results to the cognitive explanations. In the case of linking corpus data with cognitive explanations, this can be done as follows. Each observation within the corpus can be seen as an instance of individual behavior. A corpus can consequently be considered as a ‘catalogue’ of individual behavior. Within this catalogue, we can (with the corpus-based methodological framework that was set up in chapter 3) reveal patterns which are not viewable through process data but which consist of many individual decisions i.e. individual thoughts in the minds of translators (and possibly also editors) brought together. In sum, if enough translators do the same thing, a relation is established between the individual’s behavior (one translator’s behavior; one observation in the corpus) and the aggregate level (many translators’ behavior) and a pattern can be perceived. Cognitive explanations (involving the individual’s behavior) can then be used to explicate those aggregate patterns (the patterned-up behavior of many translators).

5.1.2 Linking cognitive explanations to semantic fields

In any experimental task or corpus-based study, the researcher is confronted with the lexical level as the only way to access the mental representations (and this is also the case for our study). Even in neuroimaging studies, no distinction is made between lexical and conceptual representations “because whenever a word is accessed, both its lexical and its conceptual representations are activated" (Paradis 2004, 200-201). We therefore need to clearly establish what precisely the created semantic fields represent within a cognitive explanatory framework.

In this study, we are cautious not to consider the visualized semantic fields as representations “of how knowledge or patterns of usage are actually represented in the brain” (Divjak 2010, 146)[64]. As House (2013, 51) suggested, measurements of observable

[64] Note that the same caution would have been warranted when dealing with the results of an elicitation task.

behavior (in our case corpus observations, in House's argumentation behavioral experiments) cannot really inform us about "the cognitive processes that occur in a translator's mind" nor can they "explain the nature of cognitive representations of the two languages [or] throw light on a translator's meta-linguistic and linguistic-contrastive knowledge, comprehension, transfer and reconstitution processes emerging in translation procedures" (House 2013, 50-51). To understand what exactly our measurements – contained in the semantic fields we created – can represent within a cognitive explanation (and why they do not inform us about the cognitive processes occurring in the translator's mind), we want to make a connection here with a neurolinguistic theoretical framework developed by Paradis (2004, 2007).

Paradis puts forward the idea that the neurofunctional system involved in verbal communication (the *verbal communication system*) consists of four independent subsystems which are connected to one non-linguistic conceptual level, common for all languages where concepts – "mental representations of a 'thing'" are stored (Paradis 2007, 1999)). These four subsystems are (i) implicit linguistic competence, (ii) explicit metalinguistic knowledge (iii) pragmatic ability and (iv) motivation/affect (Paradis 2004; 2007, 3). Implicit linguistic competence is acquired incidentally, stored implicitly and used automatically (Paradis 2007, 3-4). This is the level at which the model represents languages, which are considered as "neurofunctional subsystems of the language system" (Paradis 2007, 225). Lexical semantics is part of the language subsystem (but conceptual representations belong to the nonlinguistic conceptual level) (Paradis 2007, 199). Explicit metalinguistic knowledge refers to the conscious knowledge speakers have about the input to and the output from their implicit linguistic competence (but they are not conscious about the internal structure and operation of that competence) (Paradis 2007, 4). The use of metalinguistic knowledge is controlled consciously – the speaker is fully aware of the rules s/he is applying (Paradis 2004, 222). Pragmatic ability refers to the speaker's ability to infer intended meaning from the context (Paradis 2007, 4) and is important in that "pragmatic elements will determine the language to be selected for encoding and, within the language subsystems, which constructions and lexical items are most suitable to convey the intended message" (Paradis 2004, 222). Motivation or affect "is at the root of every utterance" (Paradis 2007, 5) because implicit linguistic competence as well as explicit metalinguistic knowledge are "influenced by motivation and affect during appropriation and use" (Paradis 2004, 222). Each of these four systems is "necessary, but none is sufficient for normal verbal communication" (Paradis 2007, 5), so that any kind of communicative output (for instance, a translation) is necessarily the result of all the systems working together. In this regard, each observation contained in a corpus (as well as each observation obtained via a behavioral experiment) can be seen as the cumulative result (the spoken or written communicative output) of the independent systems of the verbal communication system working together. As a consequence, we can consider our semantic fields as semantic representations of a generalization (over many translators) of these cumulative results of the systems. This implies that we do not claim our semantic fields to represent 'what happens in the mind', but rather 'what comes out of the mind' (the result rendered by the verbal communication system, the lexical items produced at the level of the language subsystem). How exactly these systems work together and whether the outcome is more (or less) due to one or another of the systems, is a neurolinguistic question we cannot possibly answer within the scope of this study. However, by considering our semantic fields as semantic representations of the output of the joint working of the systems, we can connect the cognitive explanations which we will present in the next two sections to the phenomena observed on the basis of the semantic fields presented in chapter 4.

5.1.3 Similarities and differences between the models

The two frameworks which we will present here (Halverson's GPH and Paradis' neurolinguistic theory of bilingualism) rely on the model of bilingual cognitive representation called the Revised Hierarchical Model (proposed by Kroll & Stewart 1994; see also Brysbaert & Duyck 2010; Kroll et al. 2010), which states that in the bilingual mind, there exists one non-linguistic conceptual level, common for all languages in addition to a lexical level for each of the language systems the bilingual person masters. The two models also differ in a number of respects.

First, the GPH proposes a representational model which is formulated in an attempt to answer questions of translational effects within a cognitive corpus-based translational context. The cognitive-explanational model proposed by Paradis is to be considered as a process model grounded in neurolinguistic research, but, as we will see, it is also suitable to explain translational effects on the semantic level.

Second, the GPH claims a "multicompetence perspective (Cook 2003), which emphasizes that linguistic cognition in bilinguals is *qualitatively* different from that in monolinguals" (Halverson, forthcoming, our emphasis). Paradis (2007, 22) claims that differences in representations (at the phonological, phonotactic, lexical and conceptual level) between bilinguals and monolinguals are apparently qualitative but can be accounted for by *quantitative* changes. On the conceptual level, these quantitative changes are "defined in terms of [...] number of meaningful features for concepts" (Paradis 2007, 22). For example, the presence of the conceptual features "large ball" and "small ball" in the conceptual system of the English-French bilingual make up "particular-language-driven concepts" (Paradis 2007, 23) since activation of "large ball" leads to selection of *ballon* in the French language subsystem, activation of "small ball" leads to selection of *balle* in the French language subsystem and activation of either will lead to selection of *ball* in the English language subsystem of the bilingual. Within the

English monolingual speaker's conceptual system there is no particular-language-driven concept separating "small balls" from "large balls"; the concept "ball" contains all balls, either large or small specimens. Paradis emphasizes that "[w]hat is represented may differ" but "how it is represented and processed does not" (Paradis 2007, 22). According to Paradis, the difference between unilinguals and bilinguals is thus thought to lay only in the *content* (what is represented, not how it is represented) of the representations, which may be deviant for bilinguals compared to the native speaker's norms (2007, 11). In Halverson's view, "linguistic categories in bilingual speakers [also] differ from those of monolingual speakers" (forthcoming), but these differences are not (explicitly) linked to quantitative differences.

Thirdly, the two frameworks differ in their view on the structure of linguistic categories. In Halverson's view, and following Cook (2003) and Bassetti & Cook (2011), change in the structure of linguistic categories within bilinguals happens throughout their lifetime and is a typical characteristic of bilinguals' mental representations. Paradis considers that change in structure of linguistic categories happens in monolinguals and bilinguals alike, following the same organizational principles of storage and processes:

> Under the influence of the frequent use of the other language, concepts are modified in bilinguals to include or exclude a feature or features (i.e., static interference) in the same way that concepts are modified by new experience in unilinguals (Paradis 2007, 11).

Paradis' model explicitly states that the mechanisms of mental representation (how something is represented) and of changing mental representations (change in structure of linguistic categories) work in the same way in bilinguals and unilinguals. The null-hypothesis that ensues from this, that "there is nothing in the bilingual brain that differs in nature from anything in the unilingual brain" (Paradis 2004, 189) has the advantage that no special cerebral function or mechanism(s) need to be assumed in bilinguals (Paradis 2007, 26). Since our type of research can only claim to (try to) access the contents of the representations (and not the neurological mechanisms themselves), the acceptation of this null-hypothesis is a prerequisite to apply Paradis' framework to the type of results we are dealing with.

5.2 Gravitational Pull Hypothesis

In section 2.3.5.2, we introduced Halverson's investigations as one of the most consistent bodies of research into meaning within TS. Since the beginning of the 2000s, (2003, 2010, 2013, forthcoming), Halverson has been developing a hypothesis that proposes a cognitive basis for *translation universals*, combining theoretical assumptions from Cognitive Grammar with important findings from studies of bilingualism (Brysbaert et al. 2014; Jarvis & Pavlenko 2008; Kroll & Stewart 1994, to name just a few). The cognitive grammatical model on which the GPH is based is summarized as follows by Halverson (forthcoming):

> As originally presented, the gravitational pull hypothesis assumed a cognitive grammatical model of semantic structure. In this account, all linguistic items constitute form-meaning pairings (Langacker 1987: 76), and both form and meaning are represented cognitively. Form is taken to be either graphemic or phonological, and meaning (conceptualization), in turn, is accounted for through reference to conceptual content and processes of construal (Langacker 1987: 99–146). Conceptualizations which have been used enough to become entrenched are ordered into networks of related meanings. For example, the network for a lexical item would link all of the senses of that item, and each individual sense would also be linked to synonyms (Langacker 1987: 385; Langacker 2008: 27–54).

If we now project our own visualizations within this account, we can consider each of the created semantic fields as a network for the lexical item *beginnen*, linking all of its senses (the different clusters / meaning distinctions on the semasiological level), where each individual sense (each cluster / meaning distinction) is linked to a number of synonymous lexical items (the lexemes within each cluster, the onomasiological level).

For the development of the GPH, which tries to explain the existence of *translation universals* cognitively, the following two features of these semantic networks are crucial:

> first, the relative prominence of specific elements within a network, and second, connectivity within the network, i.e. the existence and strengths of the links between network elements (Halverson, forthcoming).

The first factor that can have a certain *translational effect* is the “relative prominence of specific elements within a network”. This relative prominence is to be understood here as “the idea that some patterns of activation within schematic networks will be more prominent than others” (Halverson forthcoming) – and can be considered as *salience*. According to Halverson – and following Langacker (2008, 226) – *salience* within a schematic network can be understood as a factor of frequency of use over time (Halverson forthcoming). High frequency of use leads to entrenchment, which makes the linguistic forms (words/constructions) associated with them “more likely to be selected” (Ibid.). Originally, *gravitational pull* (Halverson 2003) was to be understood as “semasiological salience in the target language” (Halverson forthcoming). In a recent development of the GPH, Halverson distinguishes between on the one hand *salience* in the target language, which can cause the translator to be drawn towards a highly salient target language item *(magnetism)* and on the other hand *salience* in the source language, which is considered as a true form of *gravity* (or *gravitational pull*), “a cognitive force that makes it difficult for the translator to escape from the cognitive pull of highly salient representational elements in the source language” (Halverson forthcoming). On the semasiological level, *salience* can be understood as “one of a word’s many senses [being] more prominent than the others, giving it greater cognitive weight and increasing its likelihood of being selected” (Halverson forthcoming, following Geeraerts 2009, 80). *Salience* effects can also exist on the onomasiological level, where they can be detected by “look[ing] at the range of translations of a given ST item” (Halverson forthcoming). Within the GPH, *salience* is operationalized as frequency of use (Halverson, forthcoming)

The second feature is the “connectivity within the network”. The GPH also takes into account the “high frequency co-occurrence of a translation pair, either in learning or in production tasks over time, or both” (Halverson, forthcoming). Assuming that the members of a translation pair are activated together at the representational level, then, frequent activation of one member of a translation pair can strengthen the links between the members of the translation pair (Halverson, forthcoming). The so-called *connectivity*, the strength (entrenchment) of a link between two translational equivalents is also thought to potentially influence translation (Halverson, forthcoming). The three above-mentioned phenomena, salience of source language patterns, salience of target language patterns and salient translational connections could consequently cause certain characteristics to become overrepresented or underrepresented in translated language compared to non-translated language.

As for our own representations, we have seen in section 2.3.4.1 that we take into account *overlap* as an operationalization of *salience* together with frequency in order to substantiate the prototype-based nature of our visualizations. Our visualized semantic fields therefore allow us to assess the *salience* of the revealed patterns. We can thus investigate which salient source language patterns might have caused translational effects – this could be explored by looking at the SourceField of the source language of a translation. In this dissertation, such effects have been investigated in an attempt to reveal semasiological *shining through* (see section 4.6.1). Translational effects of salient target language patterns could be explored by comparing the salient patterns in translated and non-translated target language. In this study, such patterns have been investigated as semasiological and onomasiological *normalization* (see section 4.7). Finally, salient translational connections could be revealed on the basis of the joint visualization of source and target language lexemes. In this study, such an analysis was used to investigate onomasiological *shining through* (see section 4.6.2).

The cognitive explanatory concepts provided by the GPH *magnetism*, *gravitational pull* and *connectivity* can now be employed to better comprehend and explain our findings.

Onomasiological *shining through* in TransDutch$_{\mathrm{ENG}}$ (the separate clustering of *begin* and *opstarten*) could be explained as a consequence of a *connectivity*. We indeed saw (in the visualization in section 4.6.2.1) that *begin* and *opstarten* are connected to their close cognate source language lexemes. This *salient* translational connection – *connectivity* – between the source language lexeme and their Dutch target language close cognate could indeed have provoked the separate clustering of *begin* and *opstarten*. However, following this same line of reasoning, a strong *connectivity* could be claimed between *beginnen – to begin* and *starten – to start* too (*beginnen* and *starten* are also connected to their close cognate source language lexemes according to the visualization in section 4.6.2.1). However, on the basis of the GPH, we cannot explain why in this case *beginnen* and *starten* are not clustered separately.

It is more difficult to interpret semasiological *shining through* in TransDutch$_{FR}$ (the joint clustering of ACTION and SPECIFIC ACTION) as an instance of *gravitational pull*. Indeed, we cannot explain the joint clustering of ACTION and SPECIFIC ACTION in TransDutch$_{FR}$ as a consequence of the *gravitational pull* of a salient pattern (a meaning distinction in our type of analysis) in SourceFrench, because there is no such meaning distinction in SourceFrench uniting ACTION and SPECIFIC ACTION towards which the translator could have been drawn.

We concluded that the joint clustering (in TransDutch$_{ENG}$) or separate clustering (in TransDutch$_{FR}$) of ACTION and STATE AFTER ONSET could be due to either semasiological *shining through* or semasiological *normalization*. In the case of semasiological *shining through*, a salient pattern in the source language would be exerting a *gravitational pull* from which the translator could not escape. In the case of semasiological *normalization*, the translator would be attracted towards a highly salient pattern in the target language (*magnetism*). The joint clustering of ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$ and their separate clustering in TransDutch$_{FR}$ does not seem to correspond to a salient pattern that is apparent *only* in the source language or *only* in the target language (both are in fact possible). The problem is indeed that some of the changes which come about under influence of translation within the semantic fields are the consequence of very subtle influences of both the source and the target language and cannot be accounted for as a clear *pull* towards the source language or *magnetism* of the target language. The GPH can help to explain differences in patterns that are already identified as salient (in either the source or the target language) but it cannot help us determine whether a particular change in translated language is caused by a (more) subtle influence of the source language or of the target language on the translator's behavior.

Semasiological *levelling out* does not as such presuppose an influence of either source or target language, so *magnetism* or *gravitational pull* cannot be invoked to explain the phenomenon. It can, however, be tentatively explained as a consequence of *connectivity*: the visualization of the MCA of TransDutch$_{ENG}$ (see section 4.6.2.1) shows that *to start* is often translated by verbs expressing NON-LEXICALIZED INCHOATIVITY. This implies that a strong link (*connectivity*) exists between *to start* and those translational equivalents expressing NON-LEXICALIZED INCHOATIVITY. Since *to start* can be considered a central expression of inchoativity, its connectivity with a priori less central expressions of inchoativity will trigger the use of the latter, and explain why they are part of the REFERENCE CLUSTER in TransDutch$_{ENG}$. For TransDutch$_{FR}$, a similar explanation is possible: the visualization of the MCA of TransDutch$_{FR}$ (see section 4.6.2.2) shows a strong translational link between *entrer* (a central expression of inchoativity, member of the cluster with *commencer* in SourceFrench) and the verbs expressing NON-LEXICALIZED INCHOATIVITY. Again, a connectivity effect could explain the more prototypical use of the latter in TransDutch$_{FR,}$ ultimately leading to semasiological *levelling out*.

In conclusion, we tried to use the GPH here as a post-hoc interpretative framework. The explanatory concept of *connectivity* could account for onomasiological *shining through* where the connection between the source and the target language word was apparent from their joint clustering as translational pair in the HAC on the MCA of TransDutch$_{ENG}$, interpreted as a strong translational link. Although it seems indeed quite straightforward to apply this model to explain our visualizations (and, vice versa, our visualizations seem indeed to be suitable instruments to further test the GPH), our post-hoc approach has of course its limitations. The obvious disadvantage is that some of the findings which we tried to explain on the basis of the GPH cannot be understood in terms of *gravitational pull* or *magnetism* because they are not caused by *salient* patterns in the source or target language. It is indeed impossible to determine whether *gravitational pull* or *magnetism* is at play when the phenomenon under investigation (e.g. ACTION and STATE AFTER ONSET) exists similarly in both the source and the target language.

As a consequence, the GPH would better suit as an explanatory framework for cases (ideally selected beforehand) where source and target language typically reveal distinct, salient patterns. In such cases, the researcher can (more) easily determine whether a specific phenomenon in translated language can be ascribed to a *pull* towards the source language or *magnetism* of the target language.

5.3 A cognitive-explanational model from neurolinguistics

Paradis' "neurolinguistic theory of bilingualism" (2004) proposes a framework that can account for "observable data of normal behavior" as well as for behavior observed in some pathologies (Paradis 2004, 225), and is, in our view, also compatible with observable data of "translational behavior"[65]. In section 5.3.1, we will outline the main ideas behind Paradis' theory. In section 5.3.2, we will then apply the model to translation in general, before we use it as an explanatory framework for the results obtained in this study (section 5.3.3).

5.3.1 Paradis' neurolinguistic theory of bilingualism

Paradis combines three hypotheses into one theory. The "Three-Store Hypothesis" (1978; 1980; 2004, 195-203; 2007, 3-28) is based on the earlier mentioned Revised Hierarchical Model by Kroll and Stewart (1994). Originally, the Three-Store Hypothesis was formulated by Paradis as an answer to the one- or two-store hypothesis (Kolers 1968; McCormack 1977). Investigations in psycholinguistics which had made attempts to investigate "whether the two languages of bilingual speakers are represented in two memory stores or one" (Paradis 2007, 6) yielded inconsistent experimental results though. To remedy this, Paradis (1978, 1980) proposed the so-called "Three-Store Hypothesis". It states that the bilingual mind holds two separate language systems, but only one, non-linguistic cognitive system (there is convincing evidence for this from research in aphasia) (Paradis 2004, 196). This means that the (bilingual) mind disposes of a single not language-specific and non-linguistic "common conceptual system" as well as "as many subsystems as the speaker has acquired languages" (Paradis 2007, 3). The conceptual system "is ontogenetically prior and builds concepts through experience" (Paradis 2004, 198).

This hypothesis is combined with the so-called "Subsystems Hypothesis", which claims that each (language) is an independent neurofunctional subsystem, consisting of its own, independent phonology, morphology, syntax, semantics and lexicon. Each language subsystem is connected (independently of the other language subsystems) to the single conceptual system. Within the conceptual system, conceptual features are then grouped together "in accordance with the specific lexical semantic constraints of words in each language and the relevant pragmatic circumstances at the time of their use" (Paradis 2007, 3). In other words, the specific language constraints of the language subsystem will, together with the pragmatic context determine how the conceptual features will be grouped. Figure 91 schematically summarizes the components of the verbal communication system (which incorporates the two hypotheses above) consisting of one non-linguistic (language independent) conceptual level common for

[65] Paradis' theory (2004) has been proposed within cognitive TS by House (2013). Earlier work by Paradis (1994, 2000) on simultaneous interpreting has been known and applied in cognitive perspectives on simultaneous interpreting for over ten years (Christoffels 2004; Christoffels & De Groot 2005; De Groot & Christoffels 2006).

all languages and four independent (but language-dependent) subsystems: (i) implicit linguistic competence – containing semantics, morphosyntax and phonology, (ii) explicit metalinguistic knowledge (iii) pragmatic ability and (iv) motivation/affect (Paradis 2004; 2007, 3).

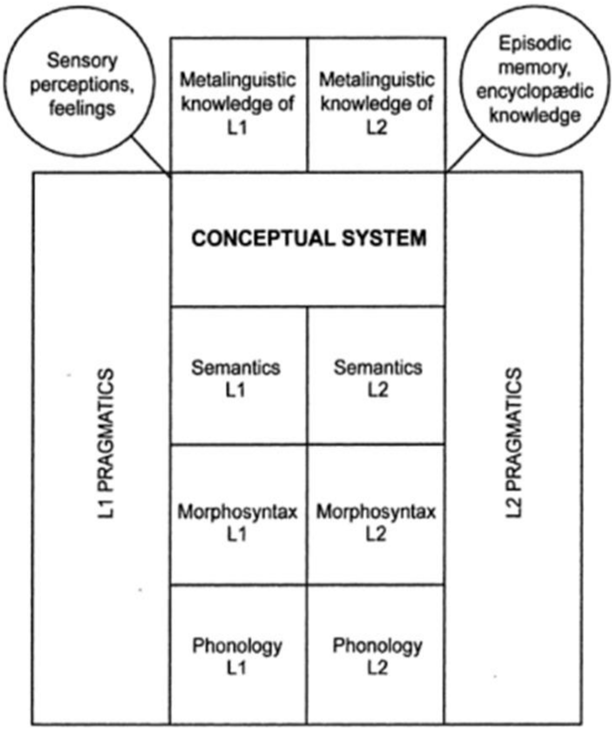

Sensory perceptions, feelings
Metalinguistic knowledge of L1
Metalinguistic knowledge of L2
Episodic memory, encyclopædic knowledge
CONCEPTUAL SYSTEM
L1 PRAGMATICS
Semantics L1
Semantics L2
Morphosyntax L1
Morphosyntax L2
Phonology L1
Phonology L2
L2 PRAGMATICS

Figure 91 Schematic representation of the components of verbal communication (copied from Paradis 2004, 227)

The selection of the appropriate conceptual features is driven by lexical meaning (Paradis 2004, 203), implying that when a speaker hears a word, the appropriate lexical item is immediately selected. The fact that the speaker is a unilingual or a bilingual does not change anything to the fact that each word is "directly perceived as a word and its meaning" (Paradis 2004, 203) (the fact that the bilingual perceives that the word is an English or a French word is of no importance to access the lexical item since such knowledge is metalinguistic in nature). This idea is captured as the "Direct Access Hypothesis", which is also compatible with the previous two hypotheses (the idea of Direct Access can be combined with the idea that the verbal communication system consists of one non-linguistic conceptual level and four independent, but language-dependent subsystems). According to the Direct Access Hypothesis "[l]exical access is language nonselective but sensitive to language-specific characteristics of the input" (Paradis 2004, 205). In other words, the lexical item that will be accessed will be the one corresponding to the perceived lexical item in the particular input language, but the language as such does not influence the accessing of the lexical item. This means that bilinguals use the available information (phonological if spoken or orthographic if written) provided by a lexical item to access the item in the according subsystem, not the meta-linguistic knowledge about 'which language the word pertains to'.

Within this hypothesis, translation equivalents are thought to function just as synonyms in a unilingual context (in cross-linguistic priming experiments, translation equivalents are predicted to cause a similar effect as synonyms (Paradis 2004, 219)), and, in general, it is stated that "when a word is activated, its synonym, homophone or translation equivalent should also receive some activation" (Paradis 2004, 219). Special attention is given to cognates, which, according to the Direct Access Hypothesis, will be immediately understood "when word forms sufficiently resemble their translation equivalent [...]" (Paradis 2004, 218). In fact, when a language user knows a word in one language as well as its cognate in another language, both language subsystems will recognize the word ("directly in one, and by immediate "completion" in the other" (Paradis 2004, 218). In cross-linguistic priming experiments, the fact that no extra processing time is needed is understood as "simultaneous activation of two languages" (Paradis 2004, 219). Simultaneous activation (no extra processing time) then reflects "either (1) the similarity of lexical meaning between a word and its translation equivalent at the conceptual level, or (2) the fact that any extra processing time for the recognition of a cognate in the other subsystem is insignificant" (Paradis 2004, 219). Consequently, simultaneous activation of two languages will be at its strongest for written cognates, where there is maximal semantic overlap (similarity of lexical meaning) and form overlap (typical for cognates) (Paradis 2004, 219).

5.3.2 Applying Paradis' theory to translation

Different from the bilingual speaker's case, the situation of "simultaneous activation of the two languages" can be assumed to be the normal cognitive state of a translator when he is carrying out a translation task, so that words with identical lexical meaning and their translations will be 'automatically' activated simultaneously (this would then be the case for close cognates as well as for 'entrenched' translation equivalent pairs).

The presence of a single conceptual system "does not imply that the same concept corresponds to a lexical item in L_x and its lexical equivalent L_z but [implies] that they share some of the same conceptual features, though each may also (and most often does) contain features not included in the other (Paradis 1978, 1997; Kroll & de Groot 1997; Costa et al. 2000)" (Paradis 2004, 198). As a consequence, translation equivalents have overlapping, but never identical conceptual representations (Paradis 2007, 12). For instance, French *cheveu* [hair growing on human scalps] and *poil* [any other hair] and Dutch *haar* [hair] (example adapted from Paradis 2004, 201) refer to what Paradis calls the same linguistic concept, but their conceptual representation will differ. The conceptual representation is that part of the linguistic concept which is activated and which consists of "only those relevant features of the linguistic concept [...] as restricted by the situation and the linguistic context in which the word is uttered" (Paradis 2007, 12). The conceptual representation of French *cheveu* in the sentence *la fille a de longs cheveux* [the girl has long hair] will be different (other features will be activated) from the conceptual representation of Dutch *haar* in the sentence *de hond heeft lang haar* [the dog has long hair] although *haar* and *cheveu* belong to the same linguistic concept. Both *cheveu* and *poil* can be translation equivalents of Dutch *haar*, but *cheveu*, *poil* and *haar* do not share all of their conceptual features, although they have many overlapping features (in fact, Dutch *haar* encompasses the features of both *cheveu* and *poil*). In Paradis' hypothesis, although the language systems (the subsystems) are independent, conceptual meanings group together conceptual features on the non-linguistic conceptual level. For *cheveu*, *poil* and *haar,* their sets of features will then overlap (2007, 13) on the non-linguistic conceptual level without being identical. The activation of differential sets of conceptual features works in the same way for unilingual synonyms such as *cheveu* and *poil* as for translation equivalents such as *cheveu* and *haar* or *poil* and *haar* (Paradis 2007, 14).

Applied to the case of the bilingual translator who needs to translate Dutch *haar* into French, the following situation arises: the translator, who is constantly primed by the source language, first enters a phase of comprehension. The written form *haar* activates the lexical item *haar* and its meaning on the subsystem level of the Dutch language. A connection is made with the conceptual level, where the lexical item *haar* causes a number of conceptual features to group together according to the specific lexical semantic constraints of *haar* in Dutch as well as according to the pragmatic circumstances evoked by the context *haar* was encountered in. Consider the following Figure 92 to be a (simplified) representation of the conceptual features activated by *haar*.

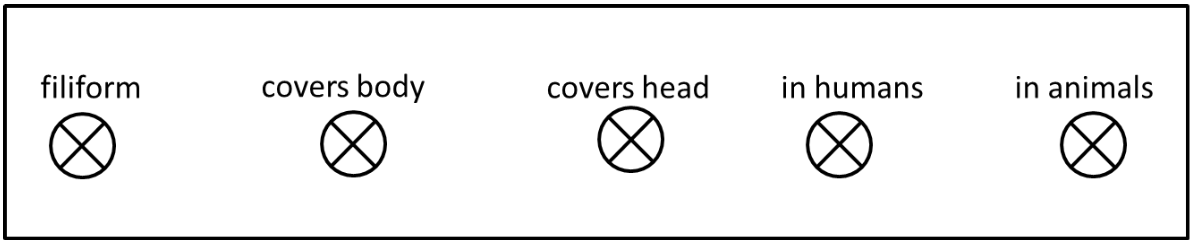


Figure 92 Representation of the conceptual features activated by *haar*

Depending on the context in which *haar* was encountered, some of the features will be activated, and others not. For the sentence *het meisje heeft lang haar* [the girl has long hair], the following conceptual features (Figure 93) will be activated (the fact that the conceptual features 'covers head' and 'in humans' are simultaneously activated, de-activates the conceptual features 'covers body' and 'in animals' for *haar* in this context):

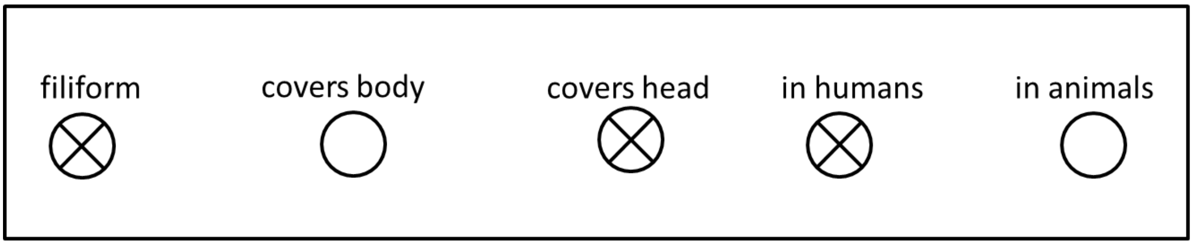


Figure 93 Activated conceptual features for *het meisje heeft lang haar*

For the sentence *de hond heeft lang haar* [the dog has long hair], the following conceptual features (Figure 94) will be activated (the activation of 'covers body' and 'in animals' de-activates 'in humans' in this context):

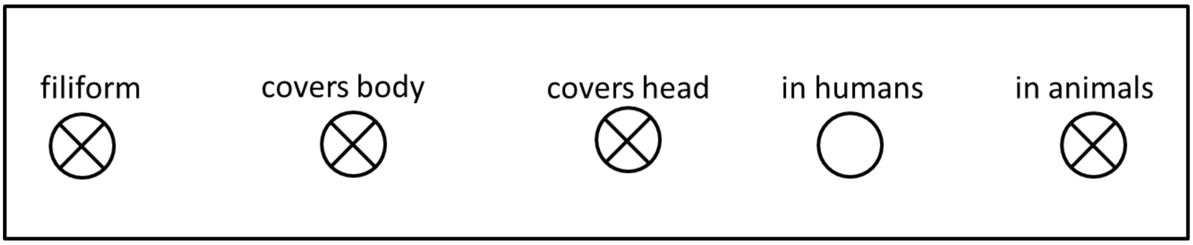


Figure 94 Activated conceptual features for *de hond heeft lang haar*

When the translator now needs to translate these two sentences with *haar* into French, s/he departs from a mental representation already activated by the lexical semantic constraints of the Dutch source language on the basis of which s/he needs to select a realization of this set (or the closest approximation to this set) of conceptual features in the target language (a lexical item in French). For the first sentence, the activated conceptual features 'filiform', 'covers head' and 'in humans' can only lead to the selection of French *cheveu* in the subsystem of the target language (since 'covers head' is not activated in *poil*). For the translation of the second sentence, however, the activated conceptual features by *haar* can lead to either *cheveu* or *poil* (the activation of the conceptual feature 'covers head' could lead to the selection of *cheveu*, but the activation of 'covers body' and 'in animals' would lead to *poil*). The conceptual features activated by *cheveu* as well as by *poil* show some overlap (but are not identical) with those activated by *haar*, as the following two Figures 95 and 96 show:

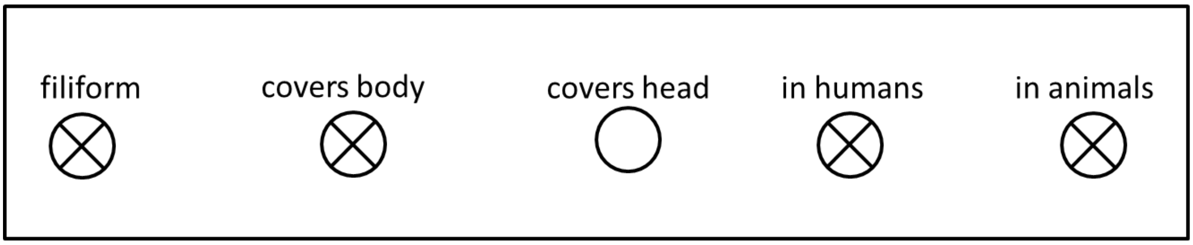


Figure 95 Activated conceptual features for *poil*

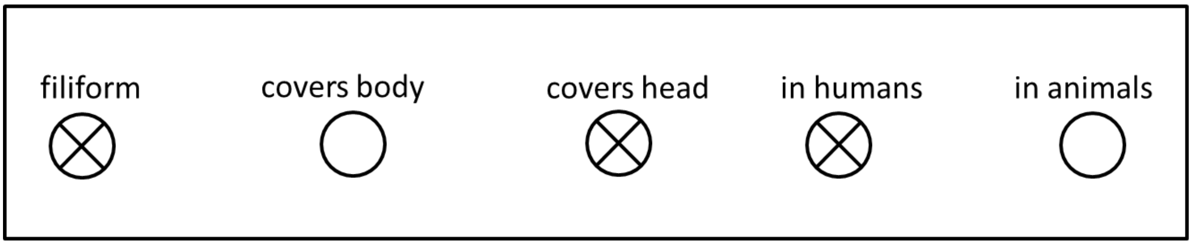


Figure 96 Activated conceptual features for *cheveu*

When the translator wants to attain sufficient overlap of conceptual features for the second sentence, the constraint on *cheveu* which does not have the conceptual feature ‘in animals’, will prevent the translator from selecting *cheveu* (since the subject of the sentence that needs to be translated refers to an animal). The activation of the conceptual feature ‘in animals’ will then prevail and lead to the selection of *poil*.

Second, when confronted with a sentence such as “de actrice is mooi” [the actress is beautiful], Dutch *actrice* activates the lexical item *actrice* and its meaning in the Dutch language subsystem (just as for *haar*), but, due to the (quasi-)total form- and meaning overlap, the French lexical item *actrice* and its meaning are simultaneously activated in the French language subsystem (and a translation is immediately found and can be produced), so that the conceptual system is not used here.

In sum, when a translator carries out a translation task, two scenarios are imaginable. First, the translator’s mind can function from the source language subsystem and arrive, via the common 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) (the example of *haar*). When the translator translates via the conceptual system, the bilingual mind first connects the lexical item (verbalized in the source language) to its appropriate concept at the common conceptual level, where the appropriate conceptual features are activated, taking into account the constraints of the source language. Then, crucially, the translator needs somehow to get rid of the constraints which the source language imposes on the concept – s/he needs to consider the nonlinguistic, unconstrained concept – and subsequently select the conceptual features which correspond to the constraints of the target language – in order to be able to select the adequate lexical equivalent in the target language (which shares some of the same conceptual features but not necessarily all features with the source language lexical item). This is where the decoding takes place; and the decision of the translator will eventually generate the production of a translation (or an omission). The translator will thus choose the lexical equivalent which shares a sufficient amount of conceptual features, comply with the constraints of the target language and consider all other constraints that can possibly act upon this choice (cultural, grammatical, pragmatic, etc.). This first ‘strategy’ in fact also explains how lack of *exact* equivalence can be bypassed by the bilingual mind (of the translator).

In the second scenario, due to the considerable form- and/or semantic overlap between the source language word and a given target language word (a cognate), the 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; 2016) (the case of *actrice*).

The importance of form similarity as put forward here is further substantiated by Brysbaert et al. (2014, 140). Although in general bilingual speaker's context "association strengths between L1 and L2 words will be very weak", they can be strong in the following three cases: for direct translations, for cognates and for so called "loan-words" (when there is no counterpart in the other language) (Brysbaert et al. 2014, 141). In a translational context with French, English and Dutch, these three cases are certainly not rare, and translators will – in all likelihood – be drawn to the selection of those direct translations, cognates and loan-words in order to translate as "quick and accurately" as possible (Kroll & Stewart 1994).

In addition to the case of cognates, where Paradis hypothesizes *direct transcoding*, it is very likely that the quick (and accurate) selection of the target language lexical item will take place for lexical items which have a *direct translation* ('entrenched translational pairs'). Although this direct translation is not a cognate, the quasi-total overlap of conceptual features and/or the association strength (what Halverson called *connectivity*) between the source language lexical item and the target language lexical item will favor the fast selection of that particular target-language lexical item. As for loan-words, the translator will become aware that the conceptual features activated by the source language lexical item correspond to extremely few or no conceptual features connected to a verbalization in the target language. Especially when none of the conceptual features are connected to a target language lexical item, the translator can choose to use the exact source language lexical item in the target language. The influence of the strong cross-linguistic associative links of direct translations, cognates and loan-words (and the degree to which these three phenomena exist within a given language pair) can possibly influence the overall translational mechanisms that are applied. In other words, although the translator might 'benefit' from language similarity (he can process translations 'quicker and more accurately'), form-similarity is likely to have a more prominent influence on the overall semantic representation of translated language when the source language is (lexically) more form-similar to the target language, because the translator seems to rely more on form-similarity (*direct transcoding*) and less on his conceptual understanding of the meaning of the unit that needs to be translated. *Translating via the conceptual system* would thus bring translators 'closer' to the (original) target language semantic representation, though never completely.

5.3.3 Applying Paradis' theory to the resulting semantic representations of inchoativity

We will now try to apply Paradis' framework to our observations about overall semasiological *levelling out* in translated language; onomasiological *shining through* in TransDutch$_{ENG,}$ semasiological *shining through* or *normalization* for ACTION and STATE AFTER ONSET in translated language and semasiological *shining through* in TransDutch$_{FR}$ for ACTION and SPECIFIC ACTION. As we mentioned in section 5.3.1, we consider our visualizations as "semantic representations of what comes out of the mind – the output of the verbal communication system". The cluster formation in each dendrogram is based on (translational and semantic) overlap and (translational co-occurrence) frequency. From the above section, we know that *direct transcoding* can only take place when a number of conditions with respect to semantic and form overlap are fulfilled. As a consequence, it seems plausible that the clustering of lexemes (especially the visualizations such as the ones presented in section 4.6.2 which jointly represent source and target language lexemes) can give indications of *direct transcoding* or *translation via the conceptual system*.

The idea of *direct transcoding* can offer a straightforward explanation for the instances of onomasiological *shining through* in TransDutch$_{ENG.}$. When the translator is working from English into Dutch, *direct transcoding* is more likely to take place since cognates between English and Dutch are much more frequent than between French and Dutch. This is confirmed by Schepens et al. (2013) who calculated the 'relative cognate frequency' (based on frequency, orthographic and phonetic similarity) for a number of language pairs and found that cognate frequency relative to translation equivalent frequency was much higher for English-Dutch (0.94, meaning that cognates have almost equal frequency of translation equivalents) than for French-Dutch (0.56, meaning that cognates have only little more than half the frequency of translation equivalents) (Schepens et al. 2013, 4). The separate clustering of *opstarten* and *begin* could be indicative that *direct transcoding* is taking place in TransDutch$_{ENG}$. However, the frequency matrix in appendix 3 shows that *opstarten* and *begin* are also translations of other lexical items, implying that there is also translation via the conceptual system taking place (although the translation of a lexeme by its close cognates does not exclude *translation by the conceptual system* of course; but for close cognates *direct transcoding* is more plausible). In contrast to *opstarten* and *begin*, and despite the fact that they also have a close cognate in English, *starten* and *beginnen*, are not forming separate (singleton) clusters. This could indicate that *direct transcoding* is taking place to a lesser extent for these two items than for *opstarten* and *begin*. No *direct transcoding* could be hypothesized for TransDutch$_{FR}$ since there are simply less close cognates between French and Dutch (especially for the field of inchoativity). The translator thus necessarily relies (more prominently) on the strategy of *translating via the conceptual* *system* when translating from a language which shares fewer close cognates with the target language such as French, compared to English. This difference could now explain why we did not find instances of onomasiological *shining through* of translated Dutch from a lexically 'less cognate' language as French.

Semasiological *levelling out* in translated language (observed as the inclusion of NON-LEXICALIZED INCHOATIVITY and *eerst* within the reference clusters of both TransDutch fields) can be explained within Paradis' neurofunctional theory as follows: target language words which do not lexicalize inchoativity or *eerst* have fewer activated conceptual features when used in their inchoative sense than more specific expressions of inchoativity (in these cases, much of the inchoativity is deduced from the context in which these lexemes are used, which implies that these lexemes only activate a minimal amount of conceptual features for inchoativity). When the translator is in search of a target-language lexical item which activates 'enough' conceptual features so that sufficient overlap with the activated conceptual features of the source language lexical item is established, the selection of a target language lexical item which only activates the minimal sufficient amount of conceptual features is in fact a 'natural choice' since it constitutes a quick, accurate and 'safe' solution. This can explain why verbs which do not lexicalize inchoativity become part of the reference cluster (with the effect of semasiological *levelling out).*

With regard to semasiological *shining through* or *normalization* for ACTION and STATE AFTER ONSET in translated language, as well as semasiological *shining through* in TransDutch$_{FR}$ for ACTION and SPECIFIC ACTION, Paradis' model also offers a possible explanation here (although it must be admitted that our interpretation is speculative and constitutes only one of the many possible ways to interpret these changes in semantic structure). We will take the example of TransDutch$_{FR}$ here, where the joint clustering of ACTION and SPECIFIC ACTION as well as the separate clustering of ACTION and STATE AFTER ONSET may be interpreted as semasiological *shining through.*

When a translator needs to translate *lancer* into Dutch, a number of conceptual features are activated by *lancer* (according to the specific lexical semantic constraints imposed by the verb as well as the context it is used in). The translator needs to select a lexical item in SourceDutch which shows a sufficient amount of overlapping conceptual features with *lancer.* Next, when the translator needs to translate *se lancer*, a number of conceptual features will again be activated (just as for *lancer*). The separate clustering of *lancer* (with ACTION) and *se lancer* (with STATE AFTER ONSET) in SourceFrench indicates that the activated conceptual features by *lancer* and *se lancer* will at least differ in that *lancer* will activate (more) conceptual features relating to ACTION and *se lancer* (more) conceptual features relating to STATE AFTER ONSET. The fact that in TransDutch$_{FR}$, ACTION and SPECIFIC ACTION are clustered together, shows that the set of common conceptual features that are maintained when translating *lancer* or *se lancer* into Dutch share a (large) amount of the common conceptual features of ACTION and SPECIFIC

ACTION, to the point that the conceptual features which usually (in non-translated language) distinguish ACTION from SPECIFIC ACTION are not activated any more, provoking the joint clustering of ACTION and SPECIFIC ACTION in TransDutch$_{FR}$. By the same mechanism, conceptual features of STATE AFTER ONSET (which are activated by *lancer*) will be de-activated because the 'pragmatic circumstances' will impose activation of conceptual features that are common to ACTION and SPECIFIC ACTION but not to STATE AFTER ONSET, provoking simultaneously also the separate clustering of ACTION and STATE AFTER ONSET. The translator's search for an adequate set of overlapping conceptual features corresponding to a lexical item in the target language subsystem can explain the joint clustering of ACTION and SPECIFIC ACTION as well as the separate clustering of ACTION and STATE AFTER ONSET in translated language.

In sum, the idea of *direct transcoding* and *translation via the conceptual system* opens a number of possibilities to explain the differences in semantic structures between translated and non-translated language. However, our interpretation suffers from the same limitations as that of the GPH in that a post-hoc application of such a framework can only go as far as adding an explanatory layer to the observations (it cannot 'test' the models as such).

5.4 Conclusion

In this chapter, we have made an attempt to explain the main observations of this study on the basis of two cognitively inspired frameworks. We first explored how the GPH could account for our results. We found that the idea of *connectivity* can explain the observed onomasiological *shining through* in TransDutch$_{ENG}$ as well as semasiological *levelling out*. Given our post-hoc approach, it appeared however difficult to connect our remaining results to the explanatory framework of the GPH since the revealed differences between the fields of translated and non-translated Dutch were not often connected to salient patterns in neither the source nor the target language.

The second cognitive framework we explored was Paradis' neurolinguistic theory of bilingualism. Onomasiological *shining through* could be explained as *direct transcoding* (which shares the basic idea with *connectivity* of salient translational relationships). Semasiological *levelling out* and semasiological *shining through* could interpreted within the wider framework as *translation via the conceptual system*.

The proposed cognitive frameworks have supplied supplementary insights into the structure of the semantic fields and in addition helped to explain where instances of *levelling out* and *shining through* on the semantic level might originate. As we already mentioned, a post-hoc application of these frameworks has its obvious limitations.

Nevertheless, we hope to have demonstrated the explanatory power of these frameworks, especially when they are combined with methodological instruments such as the visualizations proposed in our study. Much more research is nevertheless needed, so that clear hypotheses about semantic changes in translation can be drawn up *a priori* and subsequently submitted to these types of frameworks.

Chapter 6
Conclusion

Impelled by the lack of empirical studies involved with meaning variation in translation, this dissertation has placed the study of semantic differences in translation compared to non-translation at the center of its concerns. To date, much research in CBTS has focused on lexical and grammatical phenomena in an attempt to reveal presumed general tendencies of translation. On the semantic level, these general tendencies have rarely been investigated. Therefore, the goal of this study was to explore whether universal tendencies of translation also exist on the semantic level, thereby connecting the framework of *translation universals* to semantics.

Before we could set out to answer this question, a number of difficulties immediately arose: which universal tendencies were most likely to be suitable for semantic research and how could they be consequently linked to semantic inquiries? Given the attested lack of empirical studies of semantic phenomena in CBTS, no clear hypotheses could furthermore be drawn beforehand so that our method necessarily had to be explorative in nature.

In the theoretical chapter of this dissertation, we zoomed in on the relation between the study of *universals* and the study of meaning. We concluded that universal tendencies such as *levelling-out*, *normalization* and *shining through* seemed quite suitable for the investigation of meaning relationships in translation. CBTS however offers very few methodological guidelines as to how to do this. The first challenge was then to develop a methodological technique able to measure semantic similarity of translated and non-translated language. More specifically, we aimed to establish a way to visually explore semantic similarity on the basis of representations of translated and non-translated semantic fields of a concept under study.

In the methodological chapter of this work, we developed the Extended Semantic Mirrors Method, a bottom-up, statistical visualization method of semantic fields in both translated and non-translated language. The method consists of (i) a translation-driven retrieval method for the selection candidate-lexemes for a semantic field as well as (ii) a procedure to statistically visualize the retrieved data sets. Different types of visualizations were proposed in an attempt to visually explore the semantic fields so that *levelling out*, *shining through* and *normalization* could be investigated.

In the fourth chapter of this dissertation, we presented our results for the case of *beginnen* / inchoativity in Dutch. The core idea of investigating universal tendencies was transposed to the semantic level in each of the research questions. The presumed universal tendencies could now be investigated by comparing the visualizations of the semantic fields on different levels.

On the semasiological level, we formulated the following questions: **do the meanings expressed by *beginnen* (or does the prototype-based organization of those meanings) differ in translated language compared to non-translated language? Does this difference consist in *beginnen* having fewer different meanings implied in translated language compared to *beginnen* in non-translated language? If this is the case, we can call the phenomenon semasiological *levelling out*. If we indeed observe differences in the prototype-based organization of the meanings in translated and non-translated language, could there be a) an influence of the source language (*shining through*) on the translated language or b) will the expressed meanings in translated language conform to (the organization of) the meanings expressed in non-translated language (*normalization*)?**

We investigated *semasiological levelling out* by comparing the distance from each of the meanings (clusters) in a field to the center of the semantic space. We concluded that the meanings expressed by *beginnen* do differ in translated language compared to non-translated language. More specifically, we found evidence for semasiological *levelling out* in translated Dutch since in both TransDutch fields, some of the semasiological variation present in SourceDutch was ‘absorbed’ by the REFERENCE CLUSTER.

Semasiological *shining through* was investigated by comparing the semantic fields of TransDutch$_{FR}$ and TransDutch$_{ENG}$ to the semantic fields of the closest equivalents of *beginnen* in the source languages of TransDutch$_{ENG}$ and TransDutch$_{FR}$, viz. SourceEnglish *to begin* and SourceFrench *commencer*. Semasiological *normalization* was explored by comparing identical (or very similar) meaning distinctions in the visualizations of SourceDutch and TransDutch$_{ENG}$ and TransDutch$_{FR}$. We found that an influence of the source language (semasiological *shining through*) possibly provoked the joint clustering of ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$, the separate clustering of ACTION and STATE AFTER ONSET in TransDutch$_{FR}$ and the joint clustering of ACTION and SPECIFIC ACTION in TransDutch$_{FR}$. On the other hand, the specific clustering of ACTION and STATE AFTER ONSET in TransDutch$_{ENG}$ (into the REFERENCE CLUSTER) and in TransDutch$_{FR}$ (into separate clusters) could also be explained as different degrees of target language influence, and hence, semasiological *normalization*.

On the onomasiological level, we formulated the following questions: **Will the words expressing the concept of inchoativity (or the prototype-based organization of those**

words) differ in translated Dutch compared to non-translated Dutch? If we indeed observe differences in prototype-based organization of the lexemes within the clusters, do we rather see a) the organization of the lexemes in the source language semantic field *shine through* in the translated semantic field; or, b) will the organization of the lexemes within the clusters (meaning distinctions) in translated language tend to be more similar (*normalize*) to the organization of the lexemes within the meaning distinctions in non-translated target language?

We investigated changes in the prototype-based organization on the onomasiological level by evaluating the distance of each lexeme to the *centroid* (considered as the abstract prototype) of the cluster (the meaning distinction) it belongs to. We observed indeed that the prototype-based organization of lexemes within the different meaning distinctions differed in translated language, compared to non-translated language. Unfortunately, we could not connect our conclusions directly to the idea of onomasiological *levelling out*, since the number of lexemes in each visualization is kept stable. We did notice minimal changes in the prototype-based organization of the lexemes and found that lexemes which are near-synonyms in SourceDutch (such as *starten* and *beginnen*, *start* and *begin*, *oprichten* and *opzetten*) tend to become less near-synonymous in translated language.

Onomasiological *shining through* was investigated by visualizing the French and English source language lexemes together with the Dutch target language lexemes. We found that the distinct clustering of *opstarten* and *begin* (as such semasiological phenomena) in TransDutch$_{ENG}$ could be explained as an influence of the source language – *shining through* – on the onomasiological level.

Onomasiological *normalization* was investigated by comparing the prototype-based organization of the lexemes in each meaning distinction in translated Dutch to those present in non-translated Dutch. We found that the prototype-based organization of *oprichten* and *opzetten* in TransDutch$_{ENG}$ was showing signs of onomasiological *normalization* because of the similarity with the prototype-based organization of these lexemes in SourceDutch.

In chapter 5, we tried to explain the main results of this study on the basis of two cognitively inspired frameworks, in an attempt to understand *why* semantic fields of translated language differ from semantic fields of non-translated language. The proposed cognitive frameworks – the Gravitational Pull Hypothesis and Paradis' neurolinguistic theory of bilingualism – were applied to our results in an attempt to understand where *levelling out*, *shining through* and *normalization* on the semantic level might originate. Based on the idea of *connectivity* (a concept from the GPH) or *direct transcoding* (from Paradis' model), we accounted for the separate clustering of *begin* and *opstarten* in TransDutch$_{ENG}$ (onomasiological *shining through*). In addition, by following the reasoning behind *translating via the conceptual system* (Paradis), we could tentatively explain how the observed instances of semasiological *levelling out*, semasiological *shining through* or *normalization* had come about.

Retrospective insights

As we have mentioned before, our conclusions about tendencies of *levelling out*, *shining through* and *normalization* are based on observations of minimal changes in the prototype-based organization of clusters and lexemes. It must be admitted that they are moreover post-hoc interpretations of the rendered visualizations and as such naturally open for discussion. Especially on the onomasiological level, it appeared hard to convincingly connect these minimal observations to larger tendencies of translational behavior. This might indeed merely come to show that the semantic changes are primarily taking place on the semasiological level, rather than on the onomasiological level, although it is also possible that the applied approach is better fitted to discern tendencies on the semasiological level than on the onomasiological level. We indeed see that (the few) striking observations on the onomasiological level are the ones that cause semasiological change (such as the separate clustering of *opstarten* and *begin*). Without a doubt, the limited number of lexemes within our visualizations (and the fact that the number of lexemes is furthermore kept stable throughout all visualizations) is one of the reasons why general tendencies seemed much more difficult to account for on the onomasiological level. This brings us to an important point about the impact of methodological choices on our results.

Our interpretations of the observed phenomena in terms of general tendencies of translation are obviously heavily determined by the visualizations they rely on. These visualizations have come about as a result of a number of methodological choices which were taken primarily in the interest of the development of a viable visualization method of semantic fields in translated and non-translated language. Some of the choices undoubtedly impacted the overall appearance of the visualizations, and hence, influenced the further interpretation of the fields in terms of universal tendencies of translation, as the following examples illustrate.

- Our decision to select the same lexemes for each visualization was taken to ensure the comparability of the visualizations (see section 3.4.3) but had the effect that onomasiological *levelling out* could not be investigated as such.
- Our choice – inspired by pragmatic considerations (see section 3.5.1) – to exclude the verb pattern ‘to be+ing-form’ for further annotation.
- The observation of a frequency threshold of three observations – which was substantiated in section 3.4.3 – has impacted the number of selected lexemes. A frequency threshold of two observations would have resulted in the following 9 lexemes to be added to this list: *aangaan, aanvatten, begin-, doen, lanceren, maken, nemen, sinds, start-*. At first sight, the integration of these lexemes within the analysis would not have provoked any substantial changes to our visualizations:

doen [to do], *maken* [to make] and *nemen* [to take] seem to be good candidate-lexemes for the meaning distinction NON-LEXICALIZED INCHOATIVITY, *lanceren* [to launch] can be understood as a verb of SPECIFIC ACTION. We could expect *begin-* and *start-* (in compounds) to either form a distinct cluster, or be part of the REFERENCE CLUSTER. *Aangaan* [to enter into] and *aanvatten* [to commence] might form a distinct cluster, based on their more formal nature, or be part of a central cluster in the analysis.

- Our preference to base our method on the translational hypothesis rather than on the distributional hypothesis has obviously played a decisive role in the further visualization of the semantic fields.
- The determination of the meaning distinctions on the basis of cluster significance, and, more generally, the decision to carry out a HAC on the output of a CA, the chosen distance measure and clustering algorithm, have all been decisive in the 'shaping' of the semantic field structures.

As a result, it becomes clear that more research will be needed to verify the stability of the visualizations before we can focus on a more fine-tuned interpretation of the semantic fields. A number of the alternative methodological possibilities listed up above will need to be tested before we can pursue to a deeper level of analysis of the semantic fields. For example, the possibilities and limitations of the SMM++ would certainly need to be further explored to see whether the annotation of verb patterns such as 'to be+ing-form' is realistic within SMM++ (taking into account the expansiveness of the technique). In addition, 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. To this extent, a first comparison carried out by Vandevoorde et al. (2016) showed that the distributional and the translational method yield similar visualizations of the semantic field of inchoativity in Dutch.

As we already mentioned, we did not depart from clear-cut hypotheses to investigate the general tendencies of translation on the semantic level. Due to the lack of previous work on the subject, we were left in the dark about what *levelling out, shining through* or *normalization* would look like on the semantic level, which explains the explorative character of this study. As a result, our primary concern was to imagine how these so-called *universals* could possibly be operationalized on the semantic level. In this regard, we did not choose our case *beginnen* in function of testing one or the other universal (but rather out of pragmatic – corpus frequencies – considerations, as a 'good for all' test case). As a consequence, most of our main observations are not *clearly* illustrating the one or the other tendency of translational behavior. One of the striking differences between the translated fields and the non-translated field concerns the clustering of

ACTION and STATE AFTER ONSET. We failed to ascribe this phenomenon to either *normalization* or *shining through*. Since verbs of ACTION and verbs of STATE AFTER ONSET exist (although to different extents) in French, English and Dutch, our visualizations did not allow us to determine which influence (source or target language) was causing the changes in the semantic structures. Most possibly, ACTION / STATE AFTER ONSET is a case where there is neither a strong source language influence nor a prevailing target language influence, and both *normalization* and *shining through* (or none) are at play. Although it would have been more gratifying to expound clear cases of *shining through* and *normalization*, the reality of translational behavior is most probably often very similar to this situation of ACTION and STATE AFTER ONSET, where various influences cause subtle changes which ultimately alter translated language (when compared to non-translated language) but stay extremely difficult to tease apart and to capture.

Although the two cognitive frameworks which were subsequently applied did not miraculously enable us to differentiate between *shining through* from *normalization*, the idea of *translation via the conceptual system* (Paradis) offered a possible explanation for the observed phenomena in the translated semantic fields, without needing to tease apart source and target language influence (since the observed translational outcome is accounted for by what happens in the non-observable, non-linguistic conceptual system).

With this dissertation, we hope to have opened the way for more semantic research in TS. A number of methodological developments presented in this dissertation might constitute a first small step towards more research into semantic differences in translation and more cognitive explanations for translational behavior. We showed that, despite the difficulties to empirically investigate semantic phenomena, and despite notorious TS-related obstacles such as *equivalence*, it is possible to empirically investigate *translation universals* on the semantic level. The method that was put forward in this study as well as the idea to rely on statistical visualization to investigate semantic differences in translation might be further used and developed to explore semantic differences in translation and gain more insights into the mechanisms of translation on more a more abstract, semantic level. Further research will eventually lead to clear hypotheses about semantic changes in translation which can subsequently be submitted to the types of frameworks we now applied tentatively and post-hoc.

Bibliography

Agirre, Eneko, and Philip Glenny Edmonds. 2007 (2006). *Word Sense Disambiguation. Algorithms and Applications.* Kluwer Academic Publishers.

Aijmer, Karin, Ad Foolen, and Anne-Marie Simon-Vandenbergen. 2006. "Pragmatic Markers in Translation: A Methodological Proposal." In *Approaches to Discourse Particles*, edited by Kerstin Fischer, 101-14. Amsterdam: Elsevier.

Aijmer, Karin, and Anne-Marie Simon-Vandenbergen. 2003. "The Discourse Particle Well and Its Equivalents in Swedish and Dutch." *Linguistics* 41.6, no. 388: 1123-62.

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, no. 10: 1781-806.

Altenberg, Bengt. 1999. "Adverbial Connectors in English and Swedish: Semantic and Lexical Correspondences." In *Out of Corpora : Studies in Honour of Stig Johansson*, edited by Hilde Hasselgård, Signe Oksefjell and Stig Johansson, 249-68. Amsterdam: Rodopi.

Altenberg, Bengt. 2007. "The Correspondence of Resultive Connectors in English." *Nordic Journal of English Studies* 6, no. 1: 1-26.

Altenberg, Bengt, and Sylviane Granger. 2002. "Recent Trends in Cross-Linguistic Lexical Studies." In *Lexis in Contrast: Corpus-Based Approaches*, edited by Bengt Altenberg and Sylviane Granger, 3-48. Amsterdam: John Benjamins.

Alves, Fabio. 2003. "Foreword. Triangulation in Process Oriented Research in Translation." In *Triangulating Translation. Perspectives in Process Oriented Research*, edited by Fabio Alves, VII-X. Amsterdam & Philadelphia: John Benjamins.

Anthony, Laurence. 2013. "A Critical Look at Software Tools in Corpus Linguistics." *Linguistic Research* 30, no. 2: 141-61.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert, and Arne Zeschel. 2010. "Cognitive Corpus Linguistics: Five Points of Debate on Current Theory and Methodology." *Corpora* 5, no. 1: 1-27.

Baayen, R. H. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baker, Mona. 1992. *In Other Words. A Coursebook on Translation.* London & New York: Routledge.

Baker, Mona. 1993. "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.

Baker, Mona. 1995. "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research." *Target* 7, no. 2: 223-43.

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*, edited by Harold Somers, 175-86. Amsterdam & Philadelphia: John Benjamins.

Baker, Mona. 1999. "The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators." *International Journal of Corpus Linguistics* 4, no. 2: 281-98.

Baker, Mona. 2004. "A Corpus-Based View of Similarity and Difference in Translation." *International Journal of Corpus Linguistics* 9, no. 2: 167-93.

Bassetti, Benedetta, and Vivian Cook. 2011. "The Second Language User." In *Language and Bilingual Cognition*, edited by Vivian Cook and Benedetta Bassetti, 143-90. New York: Psychology Press.

Becher, Victor. 2010. "Abandoning the Notion of "Translation-Inherent” Explicitation. Against a Dogma of Translation Studies." *Across Languages and Cultures* 11, no. 1: 1-28.

Bednarczyk, Anna. 1997. "Equivalence of Translation and the Associative Unit of Translation." In *Translation and Meaning, Part 4*, edited by Marcel Thelen and Barbara Lewandowska-Tomasczyk. Maastricht: Hogeschool Maastricht.

Bernardini, Silvia, and Adriano Ferraresi. 2011. "Practice, Description and Theory Come Together: Normalization or Interference in Italian Technical Translation?" *Meta* 56, no. 2: 226-46.

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

Blum-Kulka, S. 1986. "Shifts of Cohesion and Coherence in Translation." In *Interlingual and Intercultural Communication. Discourse and Cognition in Translation and Second Language Acquisition Studies*, edited by J. House and S. Blum-Kulka, 17-35. Tübingen: Gunter Narr, 1986. Reprint, Venuti, L. (ed.) 2000. The Translation Studies Reader. London: Routledge, 298-313.

Boase-Beier, Jean. 2006. *Stylistic Approaches to Translation. Translation Theories Explored*. Manchester: St Jerome Publishing.

Brysbaert, Marc, and Wouter Duyck. 2010. "Is It Time to Leave Behind the Revised Hierarchical Model of Bilingual Language Processing after Fifteen Years of Service?" *Bilingualism: Language and Cognition* 13, no. 3: 359-71.

Carl, Michael. 2010. "Triangulating Product and Process Data: Quantifying Alignment Units with Keystroke Data." *Copenhagen studies in language* 38: 225-47.

Carletta, J. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic." *Computational Linguistics* 22, no. 1: 249-54.

Catford, J.C. 1965. *A Linguistic Theory of Translation. An Essay in Applied Linguistics*. London: Oxford University Press.

Chesterman, Andrew. 2004. "What Is a Unique Item?". In *Doubts and Directions in Translation Studies*, edited by Yves Gambier, Miriam Shlesinger and Radegundis Stolze, 3-13. Amsterdam & Philadelphia: John Benjamins.

Christoffels, Ingrid K. 2004. "Cognitive Studies in Simultaneous Interpreting." Phd thesis, University of Amsterdam.

Christoffels, Ingrid K., and Annette M.B. De Groot. 2005. "Simultaneous Interpreting. A Cognitive Perspective." In *Handbook of Bilingualism. Psycholinguistic Approaches*, edited by Judith F. Kroll and Annette M.B. De Groot, 454-79. Oxford: Oxford University Press.

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

Clark, Stephen. 2015. "Vector Space Models of Lexical Meaning." In *The Handbook of Contemporary Semantic Theory*, edited by Shalom Lappin and Chris Fox, 493-522. Chichester: John Wiley & Sons.

Cook, Vivian, ed. 2003. *The Effects of the Second Language on the First*. Clevedon: Multilingual Matters.

Costa, Albert, Angels Colomé, and Alfonso Caramazza. 2000. "Lexical Access in Speech Production. The Bilingual Case." *Psicológica* 21: 403-37.

Dagan, Ido, Alon Itai, and Ulrike Schwall. 1991. "Two Languages Are More Informative Than One." In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics* 130-137. Berkeley, California.

De Clercq, Orphée, Veronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. "Using the Crowd for Readability Prediction." *Natural Language Engineering* 20: 293-325.

De Groot, Annette M.B. 1992. "Bilingual Lexical Representation. A Closer Look at Conceptual Representations." In *Orthography, Phonology, Morphology and Meaning*, edited by Ram Frost and Leonard Katz, 389-412. Amsterdam: North Holland.

De Groot, Annette M.B., and Ingrid K. Christoffels. 2006. "Language Control in Bilinguals. Monolingual Tasks and Simultaneous Interpreting." *Bilingualism: Language and Cognition* 9, no. 2: 189-201.

Deignan, Alice. 2005. *Metaphor and Corpus Linguistics*. Amsterdam: Benjamins.

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. International Journal of Translation Studies.* 24, no. 2: 203-24.

Delaere, Isabelle. 2015. "Do Translations Walk the Line? Visually Exploring Translated and Non-Translated Texts in Search of Norm Conformity." PhD Thesis, Ghent University.

Desagulier, Guillaume. 2014. "Visualizing Distances in a Set of near-Synonyms: *Rather, Quite, Fairly* and *Pretty*." In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synoymy*, edited by Dylan Glynn and Justyna Robinson, 145-78. Amsterdam & Philadelphia: John Benjamins.

Deshors, Sandra C., and Stefan Th. Gries. 2014. "A Case for the Multifactorial Assessment of Learner Language: The Uses of May and Can in French-English Interlanguage." In *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, edited by Dylan Glynn and Justyna Robinson, 179-204. Amsterdam & Philadelphia: John Benjamins.

De Sutter, Gert. 2013. "On the Inevitability of Multivariate Statistics in Corpus-Based Translation Studies: Some Whys and Hows." In *Empirical research into translation and contrastive linguistics : objectives, methodologies, types of data*. Saarbrücken.

De Sutter, Gert, Marie-Aude Lefer, and Isabelle Delaere, eds. Forthcoming. *New Ways of Analyzing Translational Behavior*. Berlin & New York: Mouton.

De Sutter, Gert, Isabelle Delaere, and Koen Plevoets. 2012. "Lexical Lectometry in Corpus-Based Translation Studies. Combining Profile-Based Correspondence Analysis and Logistic Regression Modeling." In *Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide to Descriptive Translation Research*, edited by Michael Oakes and Ji Meng, 325-45. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Divjak, Dagmar. 2006. "Ways of Intending. A Corpus-Based Cognitive Linguistic Approach to near Synonyms in Russian." In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, edited by Stefan Gries and Anatol Stefanowitsch, 19-56. Berlin & New York: Mouton de Gruyter.

Divjak, Dagmar. 2010. *Structuring the Lexicon. A Clustered Model for near-Synonymy*. Berlin: De Gruyter Mouton.

Divjak, Dagmar, and Nick Fieller. 2014. "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.

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

Divjak, Dagmar, and Stefan Gries. 2008. "Clusters in the Mind? Converging Evidence from near Synonymy in Russian." *The Mental Lexicon* 3, no. 2: 188-213.

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*, edited by B. Lewandowska-Tomasczyk and Katarzyna Dziwirek, 273-96. 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*, edited by B. Szmrecsanyi and B. Wälchli, 174-204. Berlin & Boston: De Gruyter.

DuBay, W.H. 2004. *The Principles of Readability*. Costa Mesa, CA: Impact Information.

Dyvik, Helge. 1998. "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.

Dyvik, Helge. 1999. "On the Complexity of Translation." In *Out of Corpora : Studies in Honour of Stig Johansson*, edited by Hilde Hasselgård, Signe Oksefjell and Stig Johansson, 215-30. Amsterdam: Rodopi.

Dyvik, Helge. 2004. "Translations as Semantic Mirrors. From Parallel Corpus to Wordnet." In *Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, edited by Karin Aijmer and Bengt Altenberg, 311-26. Göteborg: Rodopi.

Dyvik, Helge. 2005. "Translations as a Semantic Knowledge Source." In *The second Baltic conference on human language technologies. Proceedings*, edited by Margit Langemets and Priit Penjam, 27-38. Tallinn: Institute of Cybernetics (Tallinn University of Technology), Institute of the Estonian Language.

Dyvik, Helge. 2011. "Semantic Mirrors." http://clara.b.uib.no/files/2011/06/semmirrors.pdf.

Ebeling, Jarle, and Signe Oksefjell Ebeling. 2013. *Patterns in Contrast*. Vol. 58: Amsterdam : John Benjamins.

Eldén, Lars, Magnus Merkel, Lars Ahrenberg, and Martin Fagerlund. 2013. "Computing Semantic Clusters by Semantic Mirroring and Spectral Graph Partitioning." *Mathematics in Computer Science* 7: 293-313.

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

Evert, Stefan. 2005. "The Statistics of Word Co-Occurrences. Word Pairs and Collocations." PhD thesis, Stuttgart.

Ferraresi, Adriano, Sylvia Bernardini, Giovanni Picci, and Marco Baroni. 2010. "Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation." In *Using Corpora in Contrastive and Translation Studies*, edited by Richard Xiao, 337-59. Newcastle upon Tyne: Cambridge Scholars Publishing.

Filipovic, Rudolf. 1969. "The Choice of the Corpus for a Contrastive Analysis of Serbo-Croatian and English." *Yugoslav Serbo-Croatian – English Contrastive Project, Studies 1* 1: 37-46.

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

Förster Hegrenaes, Claudia. 2014. "Conceptual Metaphors in Translation. A Corpus-Based Study on Quantitative Differences between Translated and Non-Translated English." In *Metaphor and Intercultural Communication*, edited by Andreas Musolff, Fiona MacArthur and Giulio Pagani, 73-88. London & New York: Bloomsbury.

Frawley, W. 1984. *Translation: Literary, Linguistic and Philosophical Perspectives*. Newark: University of Delaware Press.

Geeraerts, Dirk. 1985. "Preponderantieverschillen Bij Bijna-Synoniemen." *De nieuwe taalgids* 78: 18-27.

Geeraerts, Dirk. 1988. "Where Does Prototypicality Come From?". In *Topics in Cognitive Linguistics*, edited by Brygida Rudzka-Ostyn. Amsterdam & Philadelphia: John Benjamins.

Geeraerts, Dirk. 2006 [1988]. "Where Does Prototypicality Come From?" In *Words and Other Wonders*, edited by Dirk Geeraerts, René Dirven and John Taylor, 27-47. Berlin & New York: Mouton de Gruyter.

Geeraerts, Dirk. 2006 [1989]. "Prototype Theory. Prospects and Problems of Prototype Theory." In *Cognitive Linguistics. Basic Readings*, edited by Dirk Geeraerts, 141-65. Berlin & New York: Mouton de Gruyter.

Geeraerts, Dirk. 1990. "Homonymy, Iconicity, and Prototypicality." *Belgian Journal of Linguistics* 5: 49-74.

Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk. 2013. "Lexical Semantics from Speculative Etymology to Structuralist Semantics." In *The Oxford Handbook of the History of Linguistics*, edited by Keith Allan, 555-69. Oxford: Oxford University Press.

Geeraerts, Dirk. 2016. "Sense Individuation." In *The Routledge Handbook of Semantics*, edited by Nick Riemer, 233-47. Abingdon & New York: Routledge.

Geeraerts, Dirk, Stef Grondelaers, and Peter Bakema. 1994. *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin: Mouton de Gruyter.

Gellerstam, Martin. 1986. "Translationese in Swedish Novels Translated from English." In *Translation Studies in Scandinavia. Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, edited by L. Wollin and H. Lindquist. Lund Studies in English, 88-95. Lund: CWK Gleerup.

Gellerstam, Martin. 1996. "Translations as a Source for Cross-Linguistic Studies." In *Languages in Contrast. Papers from a Symposium on Text-based Contrastive Studies*, edited by Karin Aijmer, Bengt Altenberg and Mats Johansson, 53-62. Lund: Lund University Press.

Gilquin, Gaëtanelle. 2006. "The Place of Prototypicality in Corpus Linguistics. Causation in the Hot Seat." In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, edited by Stefan Gries and Anatol Stefanowitsch, 159-91. Berlin, Heidelberg & New York: Mouton de Gruyter.

Gilquin, Gaëtanelle. 2008. "Causative *Make* and *Faire*. A Case of Mismatch." In *Current Trends in Contrastive Linguistics. Functional and Cognitive Perspectives*, edited by Maria de los Ángeles Gómez González, J. Lachlan Mackenzie and Elsa González Álvarez, 177-201. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. Amsterdam & Philadelphia: John Benjamins.

Glynn, Dylan. 2010. "Synonyme, Lexical Fields, and Grammatical Constructions. A Study in Usage-Based Cognitive Semantics." In *Cognitive Foundations of Linguistic Usage Patterns. Empirical Studies*, edited by Hans-Jörg Schmid and Susanne Handl. Berlin & New York: Walter de Gruyter.

Granger, Sylviane. 2003. "The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies." In *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies*, edited by Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson, 17-30. Amsterdam & New York: Rodopi.

Greenacre, Michael. 2006. "From Simple to Multiple Correspondence Analysis." In *Multiple Correspondence Analysis and Related Methods*, edited by Michael Greenacre and Jörg Blasius, 41-77. London: Chapman.

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

Gries, Stefan Th. 2006a. " Introduction." In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, edited by Stefan Th. Gries and Anatol Stefanowitsch, 1-17. Berlin & New York: Mouton de Gruyter.

Gries, Stefan. 2006b. "Corpus-Based Methods and Cognitive Semantics. The Many Senses of *to Run*." In *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis*, edited by Stefan Gries and Anatol Stefanowitsch, 57-99. Berlin & New York: Mouton de Gruyter.

Gries, Stefan. 2010. "Corpus Linguistics and Theoretical Linguistics. A Love-Hate Relationship? Not Necesarily…". *International Journal of Corpus Linguistics* 15, no. 3: 327-43.

Gries, Stefan. 2012. "Behavioral Profiles: A Fine-Grained and Quantitative Approach in Corpus-Based Lexical Semantics." In *Methodological and Analytic Frontiers in Lexical Research*, edited by Gonia Libben, Gary Jarema and Chris Westbury, 57-80. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan Th. 2013. *Statistics for Linguistics with R. A Practical Introduction.* 2nd ed. Berlin & Boston: De Gruyter Mouton.

Gries, Stefan, and Dagmar Divjak. 2009. "Behavioral Profiles. A Corpus-Based Approach to Cognitive Semantic Analysis." In *New Directions in Cognitive Linguistics*, edited by Vyvyan Evans and Stephanie S. Pourcel, 57-75. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan, and Naoki Otani. 2010. "Behavioral Profiles. A Corpus-Based Perspective on Synonymy and Antonymy." *ICAME Journal* 34: 121-50.

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

Halverson, Sandra. 1996. "Norwegian-English Translation and the Role of Certain Connectors." In *Translation and Meaning, Part 3*, edited by Marcel Thelen and Barbara Lewandowska-Tomasczyk. Maastricht: Hogeschool Maastricht.

Halverson, Sandra L. 1997. "The Concept of Equivalence in Translation Studies: Much Ado About Something." *Target* 9, no. 2: 207-33.

Halverson, Sandra. 2003."The Cognitive Basis of Translation Universals." *Target* 15, no. 2: 197-241.

Halverson, Sandra. 2010. "Cognitive Translation Studies: Developments in Theory and Method." In *Translation and Cognition*, edited by Gregory Shreve and Erik Angelone, 349 - 69. Amsterdam: John Benjamins.

Halverson, Sandra. 2013. "Implications of Cognitive Linguistics for Translation Studies." In *Cognitive Linguistics and Translation. Advances in Some Theoretical Models and Applications*, edited by Ana Rojo and Iraide Ibarretxe-Antuñano, 33-74. Berlin: Mouton de Gruyter.

Halverson, Sandra. Forthcoming. "Developing a Cognitive Semantic Model: Magnetism, Gravitational Pull and Questions of Data and Method." In *New Ways of Analyzing Translational Behavior*, edited by Gert De Sutter, Marie-Aude Lefer and Isabelle Delaere. Berlin & New York: Mouton.

Hansen, Gyde. 2010. "Integrative Description of Translation Processes." In *Translation and Cognition*, edited by Gregory Shreve, M. and Erik Angelone, 189-211. Amsterdam & Philadelphia: John Benjamins.

Hansen-Schirra, Silvia. 2011. "Between Normalization and Shining-Through. Specific Properties of English-German Translations and Their Influence on the Target Language." In *Multilingual Discourse Production. Diachronic and Synchronic Perspectives*, edited by Svenja Kranich, Viktor Becher, Steffen Höder and Juliane House. Hamburg Studies on Multilingualism, 133-62. Amsterdam & Philadelphia: John Benjamins.

Hansen-Schirra, Silvia, Stella Neumann, and Erich Steiner. 2012. *Cross-Linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German.* Berlin & Boston: Mouton De Gruyter.

Hansen-Schirra, Silvia, and Erich Steiner. 2012. "Towards a Typology of Translation Properties." In *Cross-Linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*, edited by Silvia Hansen-Schirra, Stella Neumann and Erich Steiner, 255-80. Berlin & Boston: De Gruyter Mouton.

Hardie, Andrew, and Tony McEnery. 2010. "On Two Traditions in Corpus Linguistics, and What They Have in Common." *International Journal of Corpus Linguistics* 15, no. 3: 384-94.

Harris, Zellig Sabbettai. 1954. "Distributional Structure." *Word* 10: 146-62.

Hermans, Theo. 1985. "Translation Studies and a New Paradigm." In *The Manipulation of Literature. Studies in Literary Translation*, edited by Theo Hermans, 7-15. London & Sydney: Croom Helm.

Hermans, Theo. 1999. "Translation in Systems. Descriptive and System-Oriented Approaches Explained." Manchester: St. Jerome.

Heylen, Kris, Dirk Speelman, and Dirk Geeraerts. 2012. "Looking at Word Meaning: An Interactive Visualization of Semantic Vector Spaces for Dutch Synsets." In *Proceedings of EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 16-26. Avignon, France: Association for Computational Linguistics.

Heylen, Kris, Jose Tummers, and Dirk Geeraerts. 2008. "Methodological Issues in Corpus-Based Cognitive Linguistics." In *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems*, edited by G. Kristiansen and R. Dirven, 99-128. Berlin: Mouton de Gruyter.

Hiligsmann, Philippe. 2015. "Learner Corpora around the World." https://www.uclouvain.be/en-cecl-lcworld.html.

House, Juliane. 2013. "Towards a New Linguistic-Cognitive Orientation in Translation Studies." *Target* 25, no. 1: 46-59.

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

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

Ide, Nancy, Tomaz Erjavec, and Dan Tufis. 2001. "Automatic Sense Tagging Using Parallel Corpora." Paper presented at the 6th NLPRS 2001, Tokyo, Japan.

Ide, Nancy, and Yorick Wilks. 2007. "Making Sense About Sense." In *Word Sense Disambiguation. Algorithms and Applications*, edited by Eneko Agirre and Philip Edmonds, 46-73. Berlin: Springer.

Ivir, Vladimir. 1969. "Contrasting Via Translation: Formal Correspondence Vs. Translation Equivalence." *Yugoslav Serbo-Croatian – English Contrastive Project, Studies* 1: 13-25.

Ivir, Vladimir. 1970. "Remarks on Contrastive Analysis and Translation." *Yugoslav Serbo-Croatian – English Contrastive Project, Studies* 2: 14-26.

Ivir, Vladimir. 1981. "Formal Correspondence vs. Translation Equivalence Revisited." *Poetics Today* 2, no. 4: 51-59.

Ivir, Vladimir. 1983. "A Translation-Based Model of Contrastive Analysis." *Jyväskylä Cross-Language Studies* 9: 171-78.

Ivir, Vladimir. 1987. "Functionalism in Contrastive Analysis and Translation Studies ". In *Functionalism in Linguistics*, edited by R. Dirven and V. Fried, 471-81. Amsterdam & Philadelphia: John Benjamins.

Ivir, Vladimir. 1989. "Translation and Back-Translation." In *Yugoslav General Linguistics*, edited by Milorad Radovanovi, 131-44. Amsterdam & Philadelphia: John Benjamins.

Jarvis, Scott, and Aneta Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. New York: Routledge.

Jensen, Kristian T.H. 2009. "Indicators of Text Complexity." In *Behind the Mind. Methods Models and Results in Translation Process Research*, edited by Susanne Göpferich, Arnt Jakobsen and Inger Mees. Copenhagen Studies in Language, 61-80. Copenhagen: Samfundslitteratur.

Jenset, Gard B., and Barbara McGillivray. 2012. "Multivariate Analyses of Affix Productivity in Translated English." In *Quantitative Methods in Translation Studies a Practical Guide to Descriptive Translation Research*, edited by Ji Meng and Michael Oakes. Amsterdam & Philadelphia: John Benjamins.

Johansson, Stig. 1998. "On the Role of Corpora and Their Uses in Cross-Linguistic Research." In *Corpora and Cross-Linguistic Research*, edited by S. Johansson and S. Oksefjell, 3-24. Amsterdam: Rodopi.

Kaufman, Leonard, and Peter Rousseeuw. 1987. *Clustering by Means of Medoids.* Amsterdam: North-Holland.

Kaufman, Leonard, and Peter Rousseeuw. 1990 (2005). *Finding Groups in Data. An Introduction to Cluster Analysis.* New York: Wiley.

Kenny, Dorothy. 1998. "Equivalence." In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker and Kirsten Malmkjaer, 77-80. London & New York: Routledge.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation. A Corpus-Based Study.* St. Jerome Publishing.

Kipp, Michael, Jean -Claude Martin, Patrizia Paggio, and Dirk Heylen, eds. 2009. *Multimodal Corpora. From Models of Natural Interaction to Systems and Applications.* Heidelberg: Springer-Verlag Berlin.

Klaudy, Kinga. 2009. "The Asymmetry Hypothesis in Translation Research." In *Translators and Their Readers. In Homage to Eugène A. Nida*, edited by R. Dimitriu and M. Shlesinger, 283-303. Brussels: Les Éditions du Hazard.

Klaudy, Kinga. 2010. "Specification and Generalisation of Meaning in Translation." In *Meaning in Translation*, edited by B. Lewandowska-Tomasczyk and M. Thelen, 81-103. Frankfurt a.M: Peter Lang.

Kolers, P.A. 1968. "Bilingualism and Information Processing." *Scientific American* 218, March: 78-86.

Koller, W. 1979. *Einführung in Die Übersetzungswissenschaft.* Heidelberg: Quelle and Meyer.

Korning Zethsen, Karen. 2008. "Corpus-Based Cognitive Semantics: Extended Units of Meaning and Their Implications for Translation Studies." *Linguistica Antverpiensia, New Series - Themes in Translation Studies* 2008, no. 7: 249-62.

Kroll, J.F., and Annette M.B. de Groot. 1997. "Lexical and Conceptual Memory in the Bilingual. Mapping Form to Meaning in Two Languages." In *Tutorials in Bilingualism. Psycholinguistics Perspectives*, edited by Annette M.B. De Groot and J.F. Kroll, 169-99. Mahwah, NJ: Lawrence Erlbaum Associates.

Kroll, Judith F., and Erika Stewart. 1994. "Category Interference in Translation and Picture Naming. Evidence for Asymmetric Connections between Bilingual Memory Representations." *Journal of Memory and Language* 33: 149-74.

Kroll, Judith F., Janet G. van Hell, Natasha Tokowicz, and David W. Green. 2010. "The Revised Hierarchical Model. A Critical Review and Assessment." *Bilingualism: Language and Cognition* 13, no. 3: 373-81.

Kruger, Haidee. 2012. "A Corpus-Based Study of the Mediation Effect in Translated and Edited Language." *Target* 24, no. 2: 355-88.

Kruger, Haidee, and Bertus van Rooy. 2012. "Register and the Features of Translated Language." *Across Languages and Cultures* 13, no. 1: 33-65.

Kussmaul, Paul. 1995. *Training the Translator.* Amsterdam: John Benjamins.

Krzeszowski, Tomasz. 1971. "Equivalence, Congruence and Deep Structure." In *Papers in Contrastive Linguistics*, edited by Gerhard Nickel, 37-48. Cambridge: Cambridge University Press.

Krzeszowski, Tomasz. 1972. "Kontrastive Generative Grammatik." In *Reader Zur Kontrastiven Linguistik*, edited by Gerhard Nickel, 75-84. Frankfurt: Athenäum Fischer Verlag.

Krzeszowski, Tomasz. 1990. *Contrasting Languages. The Scope of Contrastive Linguistics*. Berlin & New York: Mouton de Gruyter.

Lakoff, George. 1987. *Women, Fire and Dangerous Things.* Chicago, IL: The University of Chicago Press.

Lan, Li, and Grahame Bilbow. 2007. "A Corpus-Based Investigation on Bi-Directional Business Translation." In *Translation and Meaning, Part 7*, edited by Marcel Thelen and Barbara Lewandowska-Tomasczyk. Maastricht: Hogeschool Maastricht.

Land, J., H. van den Bergh, and T. Sanders. 2009. "Zwakke Teksten Voor 'Zwakke' Lezers." In *Vakwerk 5: Achtergronden Van De Nt2 Lespraktijk: Lezingen Conferentie Bvnt2*, 54-62. Amsterdam: BV NT2.

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

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar. Volume 1. Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Langacker, Ronald. 2008. *Cognitive Grammar. A Basic Introduction*. Oxford: Oxford University Press.

Laviosa-Braithwaite, Sara. 1996a. "The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation." University of Manchester.

Laviosa-Braithwaite, Sara. 1996b. "Comparable Corpora. Towards a Corpus Linguistic Methodology for the Empirical Study of Translation." In *Translation and Meaning, Part 3*, edited by Marcel Thelen and Barbara Lewandowska-Tomaszczyk. Maastricht: Rijkshogeschool Maastricht.

Laviosa, Sara. 1998. "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." *Meta* 43: 557-70.

Laviosa, Sara. 2002. *Corpus-Based Translation Studies. Theory, Findings, Applications*. Amsterdam & New York: Rodopi.

Laviosa, Sara. 2003. "Corpora and Translation Studies." In *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies*, edited by Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson, 45-54. Amsterdam: Rodopi.

Laviosa-Braithwaite, Sara. 2004. *Unit of Translation*. Routledge Encyclopedia of Translation Studies. Edited by Mona Baker and Kirsten Malmkjaer London & New York: Routledge.

Lebart, Ludovic, and Boris G. Mirkin. 1993. "Correspondence Analysis and Classification." In *Multivariate Analysis, Future Directions 2*, edited by C. Cuadras and C.R. Rao, 341-57. Amsterdam: North-Holland Elsevier Science Publishers.

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

Lee, David Y. W. 2010. "What Corpora Are Available?". In *Routledge Handbook of Corpus Linguistics*, edited by Michael McCarthy, 107-21. Florence, USA: Routledge.

Leech, Geoffrey. 1991. "The State of the Art in Corpus Linguistics." In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, edited by Karin Aijmer and Bengt Altenberg, 8-29. London: Longman.

Lefever, Els, Veronique Hoste, and Martine De Cock. 2013. "Five Languages Are Better Than One: An Attempt to Bypass the Data Acquisition Bottleneck for WSD." In *Lecture Notes in Computer Science*, edited by Alexander Gelbukh, 343-54. Berlin: Springer.

Leopold, Edda. 2007. "Models of Semantic Spaces." In *Aspects of Automatic Text Analysis*, edited by Alexander Mehler and Reinhard Köhler, 117-38. Berlin Heidelberg: Springer-Verlag.

Levý, Jiří. 1989. "Translation as a Decision Process." In *Readings in Translation Theory*, edited by Andrew Chesterman, 99-104. Finland: Oy Finn Lectura Ab.

Lewandowska-Tomasczyk, Barbara. 2002. "Translation Studies in the Year 2000: The State of the Art. Cover Text, Cognition and Corpora." In *Translation and Meaning, Part 6*, edited by Marcel Thelen and Barbara Lewandowska-Tomasczyk. Maastricht: Hogeschool Zuyd.

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

Maks, Isa, Hennie van der Vliet, Attila Görög, and Piek Vossen. 2013. "User Documentation of Cornetto LMF. Lexical Resource for Dutch." edited by VU University. Amsterdam.

Malmkjaer, Kirsten. 1997. "Punctuation in Hans Christian Andersen's Stories and in Their Translations into English." In *Nonverbal Communication and Translation*, edited by Fernando Poyatos. Philadelphia: John Benjamins.

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

Martín de León, Celia. 2013. "Who Cares If the Cat Is on the Mat? Contributions of Cognitive Models of Meaning to Translation." In *Cognitive Linguistics and Translation. Advances in Some Theoretical Models and Applications*, edited by Ana Rojo and Iraide Ibarretxe-Antuñano, 99-122. Berlin & Boston: De Gruyter Mouton.

Marque-Pucheu, Christiane. 1999. " L'inchoatif: Marqueurs Formelles Et Lexicales Et Interprétation Logique ". In *La Modalité Sous Tous Ses Aspects*, edited by Svetlana Vogeleer, 233-57. Amsterdam & Atlanta: Rodopi.

Mauranen, Anna. 2000. "Strange Strings in Translated Language. A Study on Corpora." In *Intercultural Faultlines: Research Models in Translation Studies I: Textual and Cognitive Aspects*, edited by Maeve Olohan, 119-41. Manchester: St jerome.

Mauranen, Anna. 2004. "Corpora, Universals and Interference." In Translation Universals. Do They Exist?, edited by Anna Mauranen and Pekka Kujamäki, 65-82. Amsterdam & Philadelphia: John Benjamins.

Mauranen, Anna. 2008. "Universal Tendencies in Translation." In *Incorporating Corpora: The Linguist and the Translator*, edited by G. Anderman and M. Rogers, 32-48. Clevedon: Multilingual Matters.

McCormack, P.D. 1977. "Bilingual Linguistic Memory. The Independence-Interdependence Issue Revisited." In *Bilingualism*, edited by P.A. Hornby, 57-66. New York: Academic Press.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics*. Cambridge: Cambridge University Press.

Miller, George A. 1971. "Empirical Methods in the Study of Semantics." In *Semantics: An Interdisciplinary Reader*, edited by Danny Steinberg and Leon Jakobovits, 569-85. London & New York: Cambridge University Press.

Mortier, Liesbeth. 2010. "The Semantic Field of Continuation: Periphrastic Blijven and Continuer à." *Folia Linguistica* 44, no. 2: 401-38.

Mortier, Liesbeth, and Liesbeth Degand. 2009. "Adversative Discourse Markers in Contrast: The Need for a Combined Corpus Approach." *International Journal of Corpus Linguistics* 14, no. 3: 338-66.

Muller, Philippe, and Philippe Langlais. 2011. "Comparing Distributional and Mirror Translation Similarities for Extracting Synonyms." In *Advances in Artificial Intelligence. 24th Canadian Conference on Artificial Intelligence*, edited by Cory Butz and Pawan Lingras. Berlin Heidelberg: Springer-Verlag.

Munday, Jeremy. 2009. "Issues in Translation Studies." In *The Routledge Companion to Translation Studies*, edited by Jeremy Munday, 1-19. London & New York: Routledge.

Murphy, Gregory Leo. 2004. *The Big Book of Concepts*.

Mutesayire, Martha. 2004. "Apposition Markers and Explicitation. A Corpus-Based Study." *Language Matters: Studies in the Languages of Africa* 35, no. 1: 54-69.

Nida, Eugene. 1964. *Toward a Science of Translating*. Leiden: E.J.Brill.

Nida, Eugene A, and Charles Russell Taber. 1969. *The Theory and Practice of Translation*. Leiden: Brill.

Noël, Dirk. 2003. Translations as Evidence for Semantics: An Illustration." *Linguistics* 41, no. 4: 757-85.

Nord, Christiane. 1997. *Translating as a Purposeful Activity: Functionalist Approaches Explained.* Manchester: St. Jerome.

Oakes, Michael, and Ji Meng. 2012. *Quantitative Methods in Corpus-Based Translation Studies. A Practical Guide to Descriptive Translation Research*. Amsterdam & Philadelphia: John Benjamins.

Olohan, Maeve. 2003. "How Frequent Are the Contractions? A Study of Contracted Forms in the Translational English Corpus." *Target* 15, no. 1: 59-89.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.

Olohan, Maeve, and Mona Baker. 2000. "Reporting That in Translated English: Evidence for Subconscious Processes of Explicitation?" *Across Languages and Cultures* 1, no. 2: 141-58.

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*, edited by Peter Spyns and Jan Odijk: Springer Verlag.

Oster, Ulrike, and Heike van Lawick. 2008. "Semantic Preference and Semantic Prosody: A Corpus-Based Analysis of Translation-Relevant Aspects of the Meaning of Phraseological Units." In *Translation and Meaning, Part 8*, edited by Marcel Thelen and Barbara Lewandowska-Tomaszczyk. Maastricht: Hogeschool Zuyd.

Øverås, Linn. 1998. "In Search of the Third Code. An Investigation of Norms in Literary Translation." *Meta* 43, no. 4: 557-70.

Paradis, Michel. 1978. "Bilingual Linguistics Memory. Neurolinguistic Considerations." In *Annual Meeting of the Linguistic Society of America*. Boston.

Paradis, Michel. 1980. "Language and Thought in Bilinguals." *LACUS Forum* 6: 420-31.

Paradis, Michel. 1994. "Neurolinguistic Aspects of Implicit and Explitic Memory. Implications for Bilingualism." In *Implicit and Explicit Learning of Second Languages*, edited by N. Ellis, 393-419. London: Academic Press.

Paradis, Michel. 1997. "Représentation Lexicale Et Conceptuelle Chez Les Bilingues. Deux Langues, Trois Systèmes." In *Explorations Du Lexique*, edited by J. Auger and Y. Rose, 15-27. Quebec City: CIRAL.

Paradis, Michel. 2000. "Prerequisites to a Study of Neurolinguistic Processes Involved in Simultaneous Interpreting. A Synopsis." In *Language Processing and Simultaneous Interpreting*, edited by B. Englund Dimitrova and K. Hyltenstam, 17-24. Amsterdam: Benjamins.

Paradis, Michel. 2004. *A Neurolinguistic Theory of Bilingualism*. Amsterdam & Philadelphia: John Benjamins.

Paradis, Michel. 2007. "The Neurofunctional Components of the Bilingual Cognitive System." In *Cognitive Aspects of Bilingualism*, edited by Istvan Kecskes and Liliana Albertazzi, 3-28. Dordrecht: Springer.

Partington, Alan. 1998. *Patterns and Meanings. Studies in Corpus Linguistics*. Vol. 2, Amsterdam: John Benjamins.

Paulussen, Hans, Lieve Macken, Willy Vandeweghe, and Piet Desmet. 2013. "Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French." In *Essential Speech and Language Technology for Dutch. Results by the Stevin Programme*, edited by Peter Spyns and Jan Odijk, 185-99. Heidelberg, New York, Dordrecht & London: Springer.

Pedersen, Ted. 2007. "Unsupervised Corpus-Based Methods for WSD." In *Word Sense Disambiguation. Algorithms and Applications*, edited by Eneko Agirre and Philip Glenny Edmonds: Kluwer Academic Publishers.

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

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

Priss, U., and L.J. Old. 2005."Conceptual Exploration of Semantic Mirrors." Paper presented at the Third International Conference on Formal Concept Analysis, Berlin & Heidelberg.

Pulman, Stephen G. 1983. *Word Meaning and Belief*. London: Croom Helm.

Puurtinen, Tiina. 2004. "Explicitation of Clausal Relations: A Corpus-Based Analysis of Clause Connectives in Tarnslated and Non-Translated Finnish Childrens." In *Translation Universals. Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki, 165-76. Amsterdam & Philadelphia: John Benjamins.

Pym, Anthony. 2007. "Natural and Directional Equivalence in Theories of Translation." *Target* 19, no. 2: 271-94.

Pym, Anthony. 2008. "On Toury's Laws of How Translators Translate." In *Descriptive Translation Studies and Beyond. Investigations in Honor of Gideon Toury*, edited by Anthony Pym, Miriam Shlesinger and David Simeoni, 311-28: Benjamins.

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

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

Reiss, Katarina, and Hans Vermeer. 1991. *Grundlegung Einer Allgemeinen Translationstheorie*. 2 ed. Tübingen: Niemeyer.

Rice, Sally. 1996. "Prepositional Prototypes." In *The Construal of Space in Language and Thought*, edited by Martin Pütz and René Dirven, 35-65. Berlin & New York: Mouton de Gruyter.

Risku, Hanna. 1998. *Translatorische Kompetenz. Kognitive Grundlagen Des Übersetzens Als Expertentätigkeit*. Tübingen: Stauffenburg, 1998.

Risku, Hanna. 2002. "Situatedness in Translation Studies." *Cognitive Systems Research* 3: 523-33.

Risku, Hanna. 2004. *Translationsmanagement. Interkulturelle Fachkommunikation Im Informationszeitalter*. Tübingen: Narr.

Rojo, Ana, and Iraide Ibarretxe-Antuñano. 2013. "Cognitive Linguistics and Translation Studies: Past, Present and Future." In *Cognitive Linguistics and Translation. Advances in Some Theoretical Models and Applications*, edited by Ana Rojo and Iraide Ibarretxe-Antuñano, 3-30. Berlin & Boston: Walter de Gruyter.

Rosch, Eleanor. 1973. "On the Internal Structure of Perceptual and Semantic Categories." In *Cognitive Development and the Acquisition of Language*, edited by Timothy Moore, 111-44. New York: Academic Press.

Rosch, Eleanor. 1975. "Cognitive Representations of Semantic Categories." *Journal of Experimental Psychology* 104: 192–233.

Rosch, Eleanor. 1978. " Principles of Categorization." In *Cognition and Categorization*, edited by Eleanor Rosch and Barbara B. Lloyd, 27-48. Hillsdale, NJ: Erlbaum.

Rosch, Eleanor. 1999 [1978]. "Principles of Categorization." In *Concepts. Core Readings*, edited by Eric Margolis and Stephen Laurence, 189-206.

Rosch, Eleanor, and Carolyn Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7: 573–605.

Ruette, Tom, Dirk Geeraerts, Yves Peirsman, and Dirk Speelman. 2014. "Semantic Weighting Mechanisms in Scalable Lexical Sociolectometry." In *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*, edited by B. Szmrecsanyi and B. Wälchli, 205-30. Berlin & Boston: De Gruyter.

Sandra, Dominiek, and Sally Rice. 1995. "Network Analyses of Prepositional Meaning. Mirroring Whose Mind - the Linguist's or the Language User's?" *Cognitive Linguistics* 6, no. 1: 89-130.

Schäffner, Christina. 1999. "The Concept of Norms in Translation Studies." In *Translation and Norms* edited by Christina Schäffner, 1-8. Philadelphia: Multilingual Matters.

Schepens, Job, Ton Dijkstra, Franc Grootjen, and Walter J.B. van Heuven. 2013. *Cross-Language Distributions of High Frequency and Phonetically Similar Cognates*. doi:doi:10.1371/journal.pone.0063006.

Schmid, Hans-Jörg. 1993. *Cottage and Co., Idea, Start Vs. Begin*. Tübingen: Max Niemeyer.

Schmid, Hans-Jörg. 1996. "Introspection and Computer Corpora. The Meaning and Complementation of Start and Begin." In *Proceedings of the Seventh Symposium on Lexicography*, edited by Arne Zettersten and Viggo Hjornager Pedersen, 223-39. Tübingen: Max Niemeyer Verlag.

Simon-Vandenbergen, Anne-Marie. 2013. "English Adverbs of Essence and Their Equivalents in Dutch and French." *Advances in Corpus-Based Contrastive Linguistics: Studies in Honour of Stig Johansson* 54: 83.

Simon-Vandenbergen, Anne-Marie, and Karin Aijmer. 2002-2003. "The Expectation Marker *of course* in a Cross-Linguistic Perspective." *Languages in Contrast* 4, no. 1: 13-43.

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.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1996. "The Search for Units of Meaning." *Textus* IX: 75-106.

Sinclair, John. 1997. "Corpus Evidence in Language Description." In *Teaching and Language Corpora*, 27-39: Longman London and New York.

Smith, Carlota S. 1997. *The Parameter of Aspect*. 2 ed. Dordrecht: Kluwer Academic.

Snell-Hornby, Mary. 1992. "Word against Text. Lexical Semantics and Translation Theory." In *Translation and Meaning, Part 2*, edited by Barbara Lewandowska-Tomasczyk and Marcel Thelen, 97-104. Maastricht: Rijkshogeschool Maastricht.

Spalatin, Leonardo. 1967. "Contrastive Methods." *Studia Romanica et Anglica Zagrabiensia* 23: 29-45.

Speece, Deborah L. 1994/1995. "Cluster Analysis in Perspective." *Exceptionality* 5, no. 1: 31-44.

Spyns, Peter. 2013. "Introduction." In *Essential Speech and Language Technology for Dutch. Results by the Stevin Programme*, edited by Peter Spyns and Jan Odijk, 1-17. Heidelberg, New York, Dordrecht & London: Springer.

Stefanowitsch, Anatol. 2010. "Empirical Cognitive Semantics: Some Thoughts." In *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*, edited by Dylan Glynn and Kerstin Fischer, 355-80. Berlin: De Gruyter Mouton.

Stubbs, Michael. 1996. *Text and Corpus Analysis.* Blackwell: Oxford and Cambridge Mass.

Sun, Sanjun, and Gregory Shreve, M. 2014. "Measuring Translation Difficulty. An Empirical Study." *Target* 26, no. 1: 98-127.

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

Tabakowska, Elzbieta. 1993. *Cognitive Linguistics and Poetics of Translation*. Tübingen: Gunter Narr.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Boston: Pearson Addison Wesley.

Taylor, John. 1989. *Linguistic Categorization. Prototypes in Linguistic Theory*. Oxford: Clarendon Press.

Taylor, John. 1995. *Linguistic Categorization. Prototypes in Linguistic Theory*. 2 ed. Oxford: Clarendon Press.

Taylor, John. 2003. *Linguistic Categorization. Prototypes in Linguistic Theory*. Oxford: Oxford University Press.

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

Tirkkonen-Condit, Sonja. 2004. "Unique Items. Over- or Underrepresented in Translated Language?" In *Translation Universals. Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki. Amsterdam: John Benjamins.

Tirkkonen-Condit, Sonja, and Riitta Jääskeläinen. 2000. *Tapping and Mapping the Processes of Translation and Interpreting. Outlooks on Empirical Research*. Amsterdam: Benjamins.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Vol. 6, Amsterdam & Philadelphia: John Benjamins.

Tognini-Bonelli, Elena, and John Sinclair. 2006. "Corpora." In *Encyclopedia of Language and Linguistics*, edited by Keith Brown, 216-19. Amsterdam: Elsevier.

Toury, Gideon. 1980. *In Search of a Theory of Translation.* Tel Aviv: The Porter Institute for Poetics and Semiotics.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond.* Amsterdam: John Benjamins.

Trier, Jost. 1931. *Der Deutsche Wortschatz Im Sinnbezirk Des Verstandes: Die Geschichte Eines Sprachlichen Feldes I. Von Den Anfängen Bis Zum Beginn Des 12. Jhdts.* Heidelberg: Winter.

Turney, Peter D, and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of artificial intelligence research* 37, no. 1: 141-88.

Tyler, Andrea, and Vyvyan Evans. 2003. *The Semantics of English Prepositions.* New York: Cambridge University Press.

Vinay, Jean-Paul, and Jean Darbelnet. 1958. *Stylistique Comparée Du Français Et De L'anglais.* Paris: Les Éditions Didier.

Vandepitte, Sonia, and Gert De Sutter. 2013. "Contrastive Linguistics and Translation Studies." In *Handbook of Translation Studies*, edited by Yves Gambier and Luc van Doorslaer, 36-41. Amsterdam: John Benjamins.

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*.

Vandeweghe, Willy, Sonia Vandepitte, and Marc Van de Velde. 2007. "Introduction. A Linguistic 'Re-Turn' in Translation Studies?" *Belgian Journal of Linguistics* 21, no. 1: 1-10.

Vinay, Jean-Paul, and Jean Darbelnet. 1958. *Stylistique Comparée Du Français Et De L'anglais.* Paris: Les Éditions Didier.

Vossen, Piek, Isa Maks, Roxane Segers, Hennie van der Vliet, and H 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.

Vossen, Piek, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. "Cornetto: A Lexical Semantic Database for Dutch." In *Essential Speech and Language Technology for Dutch, Results by the Stevin-Programme*, edited by Peter Spyns and Jan Odijk, 165-83: Springer.

Weaver, Warren. 1955 [1949]. "Translation. Mimeographed." In *Machine Translation of Languages: Fourteen Essays*, edited by Erwin Reifler and William N. Locke. New York & London: The Technology Press of the Massachusetts Institute of Technology, John Wiley Chapman & Hall. Reprint, Nirenburg, Sergei, Harold L. Somers and Yorick A.Wilks (eds.). 2003. Readings in Machine Translation. Cambridge, MA: The MIT Press.

Wilss, Wolfram. 1996. *Knowledge and Skills in Translator Behaviour.* Amsterdam: John Benjamins.

Wittgenstein, L. 1953. *Philosophical Investigations*. Translated by G.E.M. Anscombe. Blackwell.

Xiao, Richard. 2009. "Theory-Driven Corpus Research. Using Corpora to Inform Aspect Theory." In *Corpus Linguistics. An International Handbook*, edited by Anke Lüdeling and Merja Kytö, 987-1008. Berlin: Mouton de Gruyter.

Xiao, Richard. 2010. "How Different Is Translated Chinese from Native Chinese? A Corpus-Based Study of Translation Universals." *International Journal of Corpus Linguistics* 15, no. 1: 5-35.

Zanettin, Federico. 2013. "Corpus Methods for Descriptive Translation Studies." In *Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC 2013),* edited by Chelo Vargas-SierraProcedia - Social and Behavioral Sciences.

Appendices

Appendix 1: First T-image beginnen$_{ENG}$

SLeng * TLdu Crosstabulation

Count

		TLdu	
		beginnen	Total
SLeng	already	1	1
	as from	1	1
	aspiring	1	1
	beginning (n)	3	3
	first of all	3	3
	fundamental	1	1
	initial	1	1
	introduction	1	1
	nascent	2	2
	new	1	1
	original	1	1
	start (n)	7	7
	start-up (n)	1	1
	to adopt	1	1
	to assume	1	1
	to be rooted	1	1
	to bear	1	1
	to begin	91	91
	to come on	1	1
	to commence	2	2
	to develop	1	1
	to embark	2	2
	to emerge	1	1
	to enter	2	2
	to gain	1	1
	to go ahead	1	1
	to go into	1	1
	to kick off	1	1
	to launch	2	2
	to let it lie	1	1
	to open	5	5
	to result	1	1
	to see	1	1
	to set up	3	3
	to start	171	171
	to start off	2	2
	to start out	6	6
	to start up	5	5
	to take up	2	2
	to talk	1	1
	to try	1	1
	to undertake	2	2
	young	1	1
Total		336	336

Appendix 2: First T-image beginnen$_{FR}$

		SLdu
		beginnen
TLfr	à l'origine	1
	à partir de	4
	aborder	2
	accomplir, s'	1
	admission	1
	amorcer	1
	apparaître	1
	arriver	1
	attaquer, s'	1
	atteler, s'	2
	avant	1
	avoir lieu	1
	commencer	164
	connaître	1
	création	1
	d'abord	5
	de	1
	débloquer, se	1
	début	14
	débutant (adj)	3
	débutant (n)	4
	débuter	41
	décider	1
	déclarer	1
	déclencher	1
	démarrer	7
	devenir	2
	donner le signal	1
	durer	1
	engager	1
	engager, s'	2
	entamer	29
	entreprendre	4
	entrer	3
	être en passe	1
	faire	1
	faire, se	1
	gagner	2
	immédiatement	1
	initialement	1
	installer	1
	installer, s'	1
	jeune (adj)	1
	lancer	10
	lancer, se	11
	livrer, se	1
	manifester, se	1
	mettre	1
	mettre en oeuvre	1
	mettre, se	12
	naître	1
	novice (adj)	1
	ouvrir	4
	ouvrir, s'	1
	partir	6
	passer	2
	plonger, se	1
	point de départ	2
	premier (adj)	2
	prendre conscience	1
	prendre cours	4
	prendre effet	2
	prendre son départ	3
	prendre, se	1
	procéder	1
	recommencer	4
	refaire	1
	remonter	2
	sortir	1
	survenir	1
	tendre	1
	tourner, se	1
	trouver ses marques	1
	venir	1
	venir de	1
Total		398

Appendix 3: Inverse T-image beginnen$_{ENG}$

TLdu * SLeng Crosstabulation

Count

		SLeng									Total
		beginning	first of all	start	to begin	to open	to set up	to start	to start out	to start up	
TLdu	aanvang	0	0	3	0	0	0	2	0	0	5
	begin	32	0	26	8	0	0	4	0	0	70
	beginnen	2	1	1	141	1	3	167	3	3	322
	eerst	1	0	2	7	0	1	7	0	0	18
	gaan	0	0	0	6	0	0	12	0	1	19
	komen	0	0	0	0	2	1	4	0	0	7
	krijgen	0	0	0	4	0	0	4	0	0	8
	ontstaan	1	0	0	1	4	0	1	0	0	7
	openen	0	0	0	1	72	3	1	0	0	77
	oprichten	0	0	0	0	0	16	1	0	4	21
	opstarten	0	0	2	1	1	2	3	0	3	12
	opzetten	0	0	0	0	1	16	0	0	0	17
	start	1	0	19	0	0	0	4	0	0	24
	starten	0	0	0	5	0	0	73	0	0	78
	van start gaan	0	0	1	4	0	0	3	0	0	8
	worden	0	0	0	1	0	0	3	0	0	4
Total		37	1	54	179	81	42	289	3	11	697

Appendix 4: Inverse T-image beginnen$_{FR}$

TLdu * SLfr Crosstabulation

Count

		SLfr																	Total
		à partir de	commencer	d'abord	début	débutant	débuter	démarrer	entamer	entreprendre	entrer	lancer	lancer, se	mettre, se	ouvrir	partir	prendre cours	recommencer	
TLdu	aanvang	0	0	0	14	0	0	0	0	0	0	0	0	1	0	0	0	0	15
	begin	0	1	0	87	0	0	0	0	0	0	0	1	0	0	0	0	0	89
	beginnen	3	78	12	10	3	16	9	11	5	4	12	5	21	1	2	3	2	197
	eerst	0	5	90	2	0	0	0	0	0	0	0	0	0	1	0	0	0	98
	gaan	0	4	0	0	0	0	0	0	0	0	0	0	6	0	9	0	0	19
	komen	0	0	0	0	0	0	0	0	0	4	3	0	1	0	1	0	0	9
	krijgen	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3
	ontstaan	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	3
	openen	0	0	0	0	0	0	0	1	0	0	3	0	0	44	0	0	0	48
	oprichten	0	0	0	0	0	0	0	0	0	0	15	3	2	0	0	0	0	20
	opstarten	0	1	0	0	0	3	7	10	1	0	24	2	0	0	0	0	1	49
	opzetten	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	3
	start	0	0	0	7	0	0	1	0	0	0	1	0	0	0	0	0	0	9
	starten	0	6	1	0	0	8	6	13	1	1	12	1	0	0	0	0	1	50
	van start gaan	0	4	0	0	0	12	5	4	1	0	3	0	0	0	1	0	0	30
	worden	0	2	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	5
Total		3	103	103	120	3	39	29	40	8	12	76	12	31	48	13	3	4	647

Appendix 5: Second T-image beginnen$_{\text{ENG}}$

SLdu * TLeng Crosstabulation

Count

		TLeng									
		beginning	first of all	start	to begin	to open	to set up	to start	to start out	to start up	Total
SLdu	aanvang	4	0	8	0	0	0	3	0	0	15
	begin	102	0	48	5	0	0	12	1	0	168
	beginnen	3	3	7	89	5	3	171	6	5	292
	eerst	0	5	0	4	0	0	5	0	0	14
	gaan	0	0	1	9	0	0	57	0	0	67
	komen	0	0	0	4	0	6	9	0	1	20
	krijgen	0	0	0	1	1	0	1	0	0	3
	ontstaan	0	0	0	1	0	1	3	0	0	5
	openen	0	0	0	0	63	3	1	0	0	67
	oprichten	0	0	0	0	3	116	5	0	0	124
	opstarten	0	0	0	3	1	11	26	0	12	53
	opzetten	0	0	0	0	0	52	0	0	0	52
	start	10	0	43	3	0	0	3	0	1	60
	starten	1	0	0	11	2	9	84	0	11	118
	van start gaa	0	0	2	4	2	3	19	1	3	34
	worden	0	0	0	3	0	0	3	0	0	6
Total		120	8	109	137	77	204	402	8	33	1098

Appendix 6: Second T-image beginnen$_{FR}$

SLdu * TLfr Crosstabulation

Count

		TLfr																	
		à partir de	commencer	d'abord	début	débutant	débuter	démarrer	entamer	entreprendre	entrer	lancer	lancer, se	mettre, se	ouvrir	partir	prendre cours	recommencer	Total
SLdu	aanvang	0	0	0	17	0	2	0	5	0	1	0	0	0	0	0	3	0	28
	begin	0	3	0	270	0	3	0	0	0	0	1	0	0	0	0	0	0	277
	beginnen	4	164	5	14	7	41	7	29	4	3	10	11	12	4	6	4	4	329
	eerst	0	16	174	4	0	0	0	0	0	0	0	0	0	0	0	0	0	194
	gaan	0	6	0	0	0	1	0	0	2	3	0	2	15	0	16	0	0	45
	komen	0	4	0	0	0	3	2	0	0	62	6	0	4	1	0	0	0	82
	krijgen	0	0	0	0	0	0	1	0	0	3	0	0	1	1	0	0	0	6
	ontstaan	0	0	0	6	0	0	0	1	0	0	0	0	0	13	0	0	0	20
	openen	0	1	0	0	0	0	0	0	0	0	1	0	0	127	0	0	0	129
	oprichten	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	4
	opstarten	0	1	0	0	0	4	6	3	1	0	16	0	0	0	0	0	0	31
	opzetten	0	0	0	0	0	0	0	0	0	0	1	1	0	2	0	0	0	4
	start	0	0	0	15	0	1	0	0	0	0	1	0	0	0	1	0	0	18
	starten	2	33	1	2	3	8	21	11	0	1	8	2	0	5	3	0	0	100
	van start gaa	0	2	0	0	0	5	2	1	0	1	4	0	0	0	0	0	0	15
	worden	0	2	0	0	0	0	0	2	0	4	0	2	2	0	0	0	1	13
Total		6	232	180	328	10	68	39	53	7	78	51	18	34	153	26	7	5	1295

Appendix 7: R script

```
# Set CRAN-mirror to "Belgium (Ghent)"
# Install package "svs"
# Load package "svs"

# Read in data file for TransDutchFR or TransDutchENG:
DAT <- read.csv2(file.choose(),header=TRUE,strip.white=TRUE)

# For SourceDutch, read in 2 data files:
DAT.FR2 <- read.csv2(file.choose(),header=TRUE,strip.white=TRUE)
DAT.ENG2 <- read.csv2(file.choose(),header=TRUE,strip.white=TRUE)

# Convert into frequency tables:
TAB.1 <- table(DAT.FR2[, c(1,2) ])
TAB.2 <- table(DAT.ENG2[, c(1,2) ])

# And combine the two tables:
DAT <- as.table(cbind(TAB.1,TAB.2))

# Carry out 'fast' correspondence analysis for TransDutch:
CSP <- fast_sca(DAT[, c(1,2) ])

# Carry out 'fast' correspondence analysis for SourceDutch:
CSP <- fast_sca(DAT)

# Generate a scree plot or a cumulative scree plot
barplot(CSP$val)
barplot(cumsum(CSP$val)/sum(CSP$val))

# Choose on which of the two varieties the analysis should be focussed
# Indicate the number of dimensions
POS <- CSP$pos1[, 1:...]
# Or:
POS <- CSP$pos2[, 1:...]

# Load pvclust:
library(pvclust)

# Carry out a HAC on CA with pvclust:
CLS <-
pvclust(t(POS),method.hclust="ward",method.dist="euclidean",nboot=3000)


# Plot the dendogram:
plot(CLS,main="...",sub="...",xlab="...")


# Determine number of clusters:
rect.hclust(CLS$hclust,h=...)
#or
pvrect(CLS,alpha=0.95)

# Validate cluster solution
# Load package "cluster"
library(cluster)
# Apply partitioning around medoids pam() to output of CA, using the same
distance measure as for HAC, with n = number of clusters in the solution
PAM<-pam(POS, n, diss = FALSE, metric = "euclidean", medoids = NULL, stand
= FALSE, cluster.only = FALSE, do.swap = TRUE, pamonce = FALSE, trace.lev =
0)
plot(PAM)


# Second validation of cluster solution via K-means
# Calculate cluster centers on a list
LST <- rect.hclust(CLS$hclust,h=...)
#or
LST <- pvpick(CLS,alpha=0.95)


# Add singleton clusters with complete_pvpick() function from svs()
# for SourceDutch
LST <- complete_pvpick(LST,rownames(COM))
# for TransDutch
LST <- complete_pvpick(LST,levels(DAT[,2]))
# for MCA
LST <- complete_pvpick(LST,unlist(lapply(DAT,levels)))

# Calculate cluster centers with centers_ca() function from svs()
# for SourceDutch
CEN <- centers_ca(POS,LST,apply(COM,1,sum))
# for TransDutch
CEN <- centers_ca(POS,LST$clusters,summary(DAT[,2]))
# for MCA
CEN <- centers_ca(POS,LST$clusters,freq_ca(DAT))

# Apply K-means
KCL <- kmeans(POS,CEN)

# Validate external cluster structure with the dist_wrt() function from svs
DIS <-dist_wrt(CEN)
dotchart(DIS,xlim=c(0,max(DIS)),bg="...",main="...",xlab="...")

# Validate internal cluster structure with the dist_wrt_centers() function
from svs
#for SourceDutch
freq= apply(COM,1,sum)
#for TransDutch
freq = freq_ca(DAT[,2])
#for MCA
freq_ca(DAT)

DIS <- dist_wrt_centers(POS, KCL, freq = apply(COM,1,sum), members_only =
FALSE)
dotchart(DIS[[...]],xlim=c(0,max(DIS[[...]])),bg="...",main="...",xlab="...
")

#Apply MCA and repeat procedure as for HAC on CA
CSP <- fast_mca(DAT)
```

Appendix 8: Distances of lexemes to centroids for SourceDutch

```
DIS
[[1]]
       aanvang           begin       beginnen           eerst            gaan 
    3.46413967      3.69363445     3.38253885      4.39454617      3.45622880 
         komen         krijgen       ontstaan          openen       oprichten 
    4.48534925      3.78598542     3.46474865      4.32177245      0.02952887 
     opstarten        opzetten          start         starten van start gaan 
    2.75535998      0.06749455     3.58117345      3.19549285      3.06686816 
        worden 
    3.55533903 

[[2]]
       aanvang           begin       beginnen           eerst            gaan 
    0.55330205      0.08908994     2.19010475      3.60359130      2.32441153 
         komen         krijgen       ontstaan          openen       oprichten 
    3.89901953      2.86275180     2.21507987      3.65865122      3.63757265 
     opstarten        opzetten          start         starten van start gaan 
    2.25815302      3.70027235     0.20740218      2.23275009      2.12361039 
        worden 
    2.48650607 

[[3]]
       aanvang           begin       beginnen           eerst            gaan 
     1.6594735       2.2927647      0.1254173       3.2708235       0.2722819 
         komen         krijgen       ontstaan          openen       oprichten 
     3.4488256       2.1115304      2.0267077       3.3377771       3.2631613 
     opstarten        opzetten          start         starten van start gaan 
     0.5508145       3.3478383      2.0028870       0.1264550       0.2694401 
        worden 
     1.1834631 

[[4]]
       aanvang           begin       beginnen           eerst            gaan 
     3.2985467       3.5270707      3.1994866       4.2819591       3.2991556 
         komen         krijgen       ontstaan          openen       oprichten 
     4.5050119       3.0419090      1.3471582       0.1718314       4.1998850 
     opstarten        opzetten          start         starten van start gaan 
     3.1747593       4.2210725      3.4319375       3.1248429       3.0678892 
        worden 
     3.4027884 

[[5]]
       aanvang           begin       beginnen           eerst            gaan 
     3.4994764       3.8466311      3.4054110       4.5189627       3.1970538 
         komen         krijgen       ontstaan          openen       oprichten 
     0.1317251       1.4928849      3.7050719       4.4902810       4.3949576 
     opstarten        opzetten          start         starten van start gaan 
     3.2705318       4.4466206      3.7221581       3.3679560       3.2158486 
        worden 
     2.1810695 

[[6]]
```

```
       aanvang            begin         beginnen            eerst             gaan 
  3.418562e+00     3.636506e+00     3.240007e+00     1.110223e-16     3.386583e+00 
         komen          krijgen         ontstaan           openen       oprichten 
  4.602946e+00     3.763519e+00     3.566664e+00     4.394572e+00     4.379157e+00 
    opstarten         opzetten            start          starten van start gaan 
  3.356807e+00     4.430260e+00     3.549602e+00     3.293686e+00     3.300870e+00 
        worden 
  3.467892e+00 
```

Appendix 9: Distances of lexemes to centroids for TransDutch$_{ENG}$

```
DIS
[[1]]
       aanvang          begin       beginnen          eerst           gaan 
     4.1772428      4.2660272      3.6606035      3.4678803      3.6048433 
         komen        krijgen       ontstaan         openen      oprichten 
     3.1391087      3.7250253      3.9346210      4.3005328      0.2004520 
     opstarten       opzetten          start        starten van start gaan 
     2.5129762      0.2476172      4.5318886      3.7323929      3.6994104 
        worden 
     3.7258800 

[[2]]
       aanvang          begin       beginnen          eerst           gaan 
     0.6167167      2.2744824      2.8871317      2.4709349      2.8256940 
         komen        krijgen       ontstaan         openen      oprichten 
     2.8694553      2.9265041      3.2177434      3.8700998      4.3190150 
     opstarten       opzetten          start        starten van start gaan 
     2.3375610      4.6485414      0.1284827      2.7788342      2.4777425 
        worden 
     2.8383236 

[[3]]
       aanvang          begin       beginnen          eerst           gaan 
    2.25776243     2.53884784     0.06521312     0.51308195     0.11345029 
         komen        krijgen       ontstaan         openen      oprichten 
    1.10378816     0.11370695     1.95464441     2.96852224     3.51758612 
     opstarten       opzetten          start        starten van start gaan 
    1.53758445     3.84721417     2.96275365     0.25003812     0.37259612 
        worden 
    0.12738579 

[[4]]
       aanvang          begin       beginnen          eerst           gaan 
     1.9830519      2.5465248      1.3054293      1.0569324      1.2230535 
         komen        krijgen       ontstaan         openen      oprichten 
     0.7203757      1.3757511      1.8270922      2.5526285      2.5663876 
     opstarten       opzetten          start        starten van start gaan 
     0.4202192      2.8859388      2.5973465      1.2945116      1.2033103 
        worden 
     1.3202771 

[[5]]
       aanvang          begin       beginnen          eerst           gaan 
     3.4640847      3.5484177      2.8753627      2.8300225      2.8883564 
         komen        krijgen       ontstaan         openen      oprichten 
     1.9726145      2.8978139      1.1745826      0.1067802      4.2131591 
     opstarten       opzetten          start        starten van start gaan 
     2.8121934      4.3223569      3.8828659      2.9107599      2.8650225 
        worden 
     2.9009154 
```

```
[[6]]
       aanvang          begin       beginnen          eerst           gaan 
  2.201321e+00   2.220446e-16   2.535833e+00   2.085160e+00   2.600465e+00 
         komen        krijgen       ontstaan         openen      oprichten 
  2.690092e+00   2.561533e+00   2.542763e+00   3.644855e+00   4.170110e+00 
     opstarten       opzetten          start        starten van start gaan 
  2.554513e+00   4.394258e+00   2.310246e+00   2.678813e+00   2.266192e+00 
        worden 
  2.624651e+00 
```

Appendix 10: Distances of lexemes to centroids for TransDutch$_{FR}$

```
DIS
[[1]]
       aanvang          begin       beginnen          eerst           gaan 
    0.12955053     0.05884857     2.36301934     3.16694244     3.91766584 
         komen       krijgen       ontstaan         openen      oprichten 
    2.73282963     2.63904093     3.38754995     4.03403545     2.72209090 
     opstarten       opzetten          start        starten van start gaan 
    2.78419901     2.84781317     0.54160901     2.65295171     2.60959733 
        worden 
    2.61519730 

[[2]]
       aanvang          begin       beginnen          eerst           gaan 
     3.9675981      4.0214808      3.4961958      4.0075011      4.6097797 
         komen       krijgen       ontstaan         openen      oprichten 
     3.5973206      3.6436850      0.9492880      0.0593305      3.5496501 
     opstarten       opzetten          start        starten van start gaan 
     3.6195418      3.6528091      3.7768222      3.5682872      3.5784474 
        worden 
     3.5894073 

[[3]]
       aanvang          begin       beginnen          eerst           gaan 
     2.6680740      2.7446571      1.1318943      2.7706002      3.6749542 
         komen       krijgen       ontstaan         openen      oprichten 
     1.4586546      1.5443559      2.6444725      3.6344425      0.2037736 
     opstarten       opzetten          start        starten van start gaan 
     0.2664611      0.4267752      2.1561010      0.1160007      0.4316780 
        worden 
     1.2349533 

[[4]]
       aanvang          begin       beginnen          eerst           gaan 
     2.3641785      2.4846664      0.6576414      1.7263104      2.7439995 
         komen       krijgen       ontstaan         openen      oprichten 
     1.1315485      0.9121243      2.7598750      3.5554558      1.6826330 
     opstarten       opzetten          start        starten van start gaan 
     1.7438098      1.8862396      2.0667548      1.4000293      1.2147151 
        worden 
     0.9202239 
```